Scholars Journal of Engineering and Technology

OPEN ACCESS

Abbreviated Key Title: Sch J Eng Tech
ISSN 2347-9523 (Print) | ISSN 2321-435X (Online)
Journal homepage: https://saspublishers.com

# Advancing Natural Language Processing for Underrepresented Tibeto-Burman Languages in Northeast India

N John Kuotsu[1*]

[1]Department of Computer Science, Fazl Ali College, Mokokchung, India

DOI: https://doi.org/10.36347/sjet.2024.v12i12.001 | Received: 23.10.2024 | Accepted: 29.11.2024 | Published: 04.12.2024

*Corresponding author: N John Kuotsu
Department of Computer Science, Fazl Ali College, Mokokchung, India

## Abstract

Review Article

Natural Language Processing (NLP) plays a vital role in bridging digital divides by facilitating communication and information access across diverse linguistic communities. This paper analyses the particular difficulties and the potential directions of constructing NLP resources for the Tibeto-Burman languages that are used in the North-East India: The region is linguistically diverse; however, it does not possess large-scale language resources. Some of the special characteristics of the Tibeto-Burman languages include tonal system, morphological complexity, distinctive syntactic features and script variety which present major obstacles to NLP applications including machine translation, speech recognition and named entity identification. The lack of digital corpora and annotated datasets is another challenge which adds to the problem. Focusing on the case studies, as well as the recent trends of the field, this work describes the emerging and promising techniques like transfer learning, data augmentation, and community-driven development. These methods seek to address the issues of limited data and increase the efficiency and effectiveness of applying NLP tools to the Tubeto-Burman languages in North-East India. Ultimately, improving NLP tools in this area helps to strengthen language documentation and ensure equal access to digital resources, as well as promoting the global development of NLP field knowledge and research. Addressing these challenges can pave the way for more inclusive and effective communication technologies across diverse linguistic landscapes.

**Keywords:** Natural Language Processing (NLP), Tibeto-Burman Languages, Low-Resource Languages, Transfer Learning, Data Augmentation, Digital Inclusion.

## 1. INTRODUCTION

There are many languages of the Tibeto-Burman class and many other in North-East India. The region comprises of eight states, namely Arunachal Pradesh, Assam, Manipur, Meghalaya, Mizoram, Nagaland, Sikkim, and Tripura; the place is considered one of the most linguistic diverse area in the world [1]. Still, a great number of these languages can be viewed as low-resource in the field of Natural Language Processing (NLP), which means that they entail different challenges as well as opportunities for the researchers and developers in the field.

There are several reasons why it is important to develop the NLP tools for the low-resource languages spoken in North-East India. First, it assists in overcoming language inequality and supports the use of endangered languages [2]. Secondly, it makes it possible for the people speaking these languages to get or at least generate digital contents in their languages, thus reducing the digital gap [3]. Finally, it is useful in understanding the diversity of languages of the global community and in studying the properties of language as a universal human phenomenon [4].

The focus of this paper is to examine the current state of NLP for low-resource languages of North-East India, understand the challenges and discuss directions to overcome them. By adopting this approach to the particular region of interest the study aim at strengthening the general discourse on NLP for low-resource languages and stressing the role of language technologies for multilingual societies.

## 2. Linguistic Landscape of North-East India

North-East India is home to languages from various families, including Tibeto-Burman, Indo-Aryan, and Austroasiatic. However, Tibeto-Burman languages dominate the linguistic landscape [5]. Some of the major Tibeto-Burman languages in the region include: Bodo (Assam), Manipuri/Meitei (Manipur), Mizo (Mizoram), Kokborok (Tripura), Ao, Angami, Lotha (Nagaland), Garo (Meghalaya). These languages exhibit diverse and complex linguistic features, which present unique challenges for NLP tasks:

### 2.1 Tonal Systems

Some of the languages spoken in North-East India are tonal thus, the pitch variation determines the meaning of words. This is different from languages like English where the order of CONSONANTS and VOWELS in a word dictates the meaning of the said word. Tonal languages can have a variety of tonal inventories; the number of contrastive tones varies from two up to six or more [5]. Tones are very significant in the meaning of the words in North-East Indian languages. It can be seen that a single syllable may be associated with completely opposite meanings depending on the pitch only. For example, Mizo has a complex tonal system with rising, falling, and level tones [6].

### 2.2 Morphological Complexity:

The members of the Tibeto-Burman family of languages spoken in North-East India are especially characterized by complex morphological patterns. Morphology focuses on how the units of meaning in words which are known as morphemes are combined to from other words. This characteristic differentiates these languages from others such as English languages where most of the meaning is conveyed by the position of words. For instance, Manipuri exhibits feature of an agglutinative language in that more than one morpheme can be used to form complex words [7].

### 2.3 Syntactic Features:

Some of the primary syntactic features found in the members of the Tibeto-Burman language family include the subject-object-verb or the SOV word order. This syntactic structure is present in most of the languages in the family and has numerous impacts for natural language processing applications. In the general sense, the, common word order of most of the Tibeto-Burman languages is Subject-Object-Verb (SOV), but within this family there is much variation [8]. It is important to comprehend these features in order to design suitable NLP tools for such languages.

### 2.4 Script Diversity:

**Derived from Bengali:**

Today, several North Eastern Indian languages have borrowed scripts from the Bengali script. This entails Meitei Mayek, the script for Manipuri [9]. These scripts have got close relationship all originating from this one script but they have also slightly deviated to meet the phonetic requirements of the respective language.

**Latin alphabet Adaptations:**

Another large group of scripts can be distinguished and it belongs to the scripts derived from the Latin alphabet. Mizo (Mizoram) and Ao (Nagaland) are fine instance [10, 11]. It should be noted that these scripts usually include extra characters or other letter forms called diacritics (comprise of marks placed on letters to indicate sounds that are not used in English).

## 3. Challenges in NLP for North-East Indian Languages:

### 3.1 Data Scarcity:

Some of the key issues that arise during the utilization of NLP tools for the North-East Indian languages are the scarcity of digital corpora along with less available annotated data sets. Many of these languages have a primarily oral tradition, with limited written resources [12]. This challenge is aggravated by the limited number of these languages that are used in NLP studies, holding back advancements in relevant topics of study such as machine translation and named entity recognition.

### 3.2 Linguistic Complexity:

Tibeto-Burman languages in North-East India are noted for linguistic complexity that present challenges to the processing of computational resources on language such as in natural language processing or NLP. Languages, some of which include Mizo with tone systems affecting meaning [6], and Manipuri with complex affixation [7], demand extreme accurate models for instances like speech recognition or machine translation. Furthermore, most of these languages have complicated syntactic structures with somewhat rigid sentence structures such as the SOV and thus require the use of flexible systems to address the various linguistic structures. It is necessary to know various syntactic relations and its intricacies because the second language can be highly dimensional in terms of the complexity of language in NLP systems to process and generate text [13]. Most of the current Automatic Speech Recognition systems that are developed to address the non-tonal languages do not pick these subtle differences and this leads to huge errors.

### 3.3 Script and Orthography Issues:

As most of the Tibeto-Burman North-East Indian languages are written in different scripts, there lies the problem for text normalization or optical character recognition as there are no unified systems of writing in those languages resulting in inconsistent scripts. Character sets and handwriting styles differ between languages and make it is difficult to recognize similar characters needed in multilingual research [14].

### 3.4 Lack of NLP Tools and Resources:

As for the Tibeto-Burman languages in North-East India, the basic NLP tools including POS taggers, NER and dependency parsers are largely un-developed [15]. In resource-rich languages, these pieces of work are inevitable for text preparation and analysis, while for many North-East Indian languages, the absence of such resources is one of the major reasons of NLP application. Lack of good quality POS taggers restricts the grammatical analysis of written material whereas poor NER tools hamper the extraction of data from text. Moreover, the absence of dependency parsers hampers syntactic parsing abilities, which are needed for complex tasks such as machine translation and text understanding. Such tools are needed for the further enhancement of NLP in these languages [16].

### 3.5 Multilingualism and Code-Switching:

Bilingualism is very common among the speakers in North-East India and many of them use more than one language simultaneously to communicate, and to use media; sometimes, they switch between the native Tibeto-Burman language and a dominating regional or national language such as Hindi or English [17]. This practice becomes problematic in NLP tasks because the content can include both languages. Due to the fact that standard monolingual NLP models are not capable of correctly identifying the language and performing syntactic and semantic analysis of the text in code-switching, they are not very useful. Solving these issues calls for proper and strong multilingual processing and context-based language flipping systems [18].

## 4. Opportunities and Innovative Approaches:

### 4.1 Transfer Learning and Multilingual Models:

Recent innovation in transfer learning and multilingual models indicate a promising future for a low resource language of North-East India. The work done under AI4Bharat of developing the IndicBERT model that comprises of multiple NE-Indian languages is one such achievement [19]. This approach is based upon the fact that related languages have many similar linguistic characteristics which will enhance the performance of numerous NLP tasks. It means that the models trained on high-resource languages can actually be further trained for other Tibeto-Burman languages which may show further improvement with little data [20]. As a transfer learning method, this approach enables researchers to leverage on the abundance or excess data and pre-trained models in the major world languages to meet the specialized demands of low resource languages. Hence, applying these techniques, some of the difficulties arising from the dearth of data in North-East Indian languages can be tackled by the researchers.

### 4.2 Data Augmentation Techniques:

To address data scarcity, researchers are exploring various data augmentation techniques. These include synthetic data generation from back-translation and leveraging unlabeled data through self-supervised learning.

**Back-Translation:**

Is a valuable data augmentation technique that aids in improving the quality of machine translation systems by generating synthetic parallel data. This method involves translating monolingual text into a target language and then back into the original language, creating additional training data. By utilizing back-translation, researchers can effectively enhance translation quality, especially in low-resource settings where parallel corpora are limited [21, 22].

**Self-Supervised Learning**

Tackles this challenge by creating its own supervisory signals from unlabeled data. It essentially invents its own learning tasks and goals based on the inherent structure of the data itself. Like Predicting the next word in a sentence, reconstructing masked words, or learning word embeddings [23].

### 4.3 Community-Driven Resource Development:

Engaging language communities in resource development can help overcome the data scarcity challenge.

**Speech Data Collection:**

Platforms like Common Voice can be adapted for low resource languages. Speakers can contribute voice recordings, helping to build speech corpora essential for speech recognition and other NLP tasks [24].

**Text Annotation:**

Community members can collaborate on tasks like annotating text data for part-of-speech tagging, named entity recognition, and sentiment analysis. This can be gamified or made incentive-based to encourage participation [25].

**Establishing Language Translation Centres:**

These centres can act as repositories for documenting data, annotating and translating them, and also bring together linguists, researchers, and members of the community. This can further facilitate the development of comprehensive language resources.

### 4.4 Interdisciplinary Collaboration:

Interdisciplinary Collaboration: Incorporating concepts from linguistics, anthropology, as well as computer science would pave ways to well-implemented NLP solutions particularly to low-resource languages. For this reason, it is necessary to analyze the linguistic and cultural peculiarities of the language to develop correct and culturally relevant NLP tools. Morphological and syntactic experts can enlighten on the structure of the given languages of the group, and anthropological background may unveil information on the use of the language in distinct social scenarios. Computer scientists can then use this information to improve on the efficiency of algorithms and models that will be developed in the future. Such a methodology ensures that NLP tools do not only implement well from a technical perspective but also maintain cultural sensitivity in the language being used [26].

### 4.5 Low-Resource NLP Techniques:

Some novel approaches particularly for low resource conditions seem to provide new path ways in

NLP of the Tibeto-Burman languages of the North-East India. Specifically, few-shot learning and unsupervised machine translation are quite remarkable.

**Few-Shot Learning:**

Few-shot learning attempts to train models using limited amount of labelled data which will be helpful for languages that are low-resource and limited data sets. Few-shot learning makes it possible to learn on a limited dataset to provide relatively high accuracy in multiple NLP tasks [27].

**Unsupervised Machine Translation:**

The approach of unsupervised machine translation has the advantage of not using parallel corpora and using methods such as back- translation and iterated training on monolingual data. Such an approach has proved to have positive results in improving the translation performance rates in languages with reduced resources [28].

## 5. CASE STUDIES

### 5.1 Machine Translation for Manipuri:

Recent studies have focused on the issue of language translation from English to Manipuri which is a low-resource language of North-East India. Researchers came up with an approach in machine translation that is only partially supervised, which included concepts such as self-training as well as back translation. Since Manipuri has a rather complicated grammar with its stressed morphology and engaging in agglutinative processes they have developed a specific multi-reference testing sets. Through the manipulation of the noise incorporated in to the input data they were able to enhance the translation and had a boost of 0.9 BLEU score. They also employed transfer learning, where they fine tune other models trained on similar but more prevalent language pairs for the Manipuri-English pair. This approach proved to be much better than using a classical supervised and pre-trained model, which shows the prospects of semi-supervised and transfer learning for improving low-resource languages translation [29].

### 5.2 Speech Recognition for Ao Naga:

Some recent works on Ao Naga, a tonal language from North-East India, have shown promising improvement in Automatic Speech Recognition (ASR) and Dialect Identification (DID) even with limited resources. Deep learning techniques and data augmentation strategies used in the work enhanced the development of ASR for Ao. According to the diverse feature investigation in DID study, the best core feature to be extracted is essentially the gammatonegram of the linear prediction residual which is highly suitable to the tonal characteristics of Ao. Due to the lack of data in this scenario, researchers resorted to data augmentation approaches such as speed perturbation and adding background noise, which helped increase the speech corpus and improve the system's DID by 14 percent [30]. It shows that the integration of the specialized features with data augmentation is beneficial for the under-represented languages. In addition to increasing the knowledge about tonal language processing for Ao, the improvement of both ASR and DID contributes to the better performance of NLP in other low-resource linguistic contexts.

### 5.3 Named Entity Recognition for Bodo:

The Deep learning study investigating Named Entity Recognition (NER) in Bodo, a language spoken in Assam India also holds good example of its utilization in under-resourced languages. One of the limitations arising from the limited availability of data was effectively managed by the researchers through the use of Docanno to generate the NER-tagged dataset along with the data augmentation methods. Several deep learning models were used and compared, namely of Long Short-Term Memory (LSTM), character-based models, Gated Recurrent Units (GRU), and Convolutional Neural Networks (CNN). The above models obtained good evaluation measures of accuracy, precision, recall, and F-score of Bodo NER task. Nevertheless, the findings of this research and the effective use of deep learning in tackling a resource-limited language condition suggest the applicability of such techniques for other linguistic environments. The overview of the method that aims at creating and augmenting the datasets, as well as implementing an NER model, is useful for constructing similar systems in other low-resource languages [31].

### 6. Ethical Considerations:

Ethical concerns should be taken into consideration when creating NLP technologies for North-East Indian languages. These include privacy of the information to be collected, cultural differences, and socio-economic effects which the execution of the project might bring along. Privacy is important for the collection and processing of the linguistic data and must also take into account an individual's right to privacy and his/her consent [32]. Cultural aspect: NLP tools should also be contextually appropriate to the cultural difference of the languages that the tool was developed for, in a way that is not misleading or abusive of the cultural material. Additionally, developers need to be very sensitive and guarantee that these technologies do not support the language shift or the minority languages' exclusion. NLP tools should complement the usage of minority language instead of substituting them with a more dominant language. This involves advocating language preservation and encouraging the use of minority languages in digital spaces to maintain linguistic diversity and cultural heritage [33]. It is imperative to address these ethical considerations as these are significant for building sustainable and fair NLP applications.

### 7. Future Directions:

Looking ahead, several promising directions emerge for NLP in North-East Indian languages:

### 7.1 Multilingual Pre-Training:

Developing pre-trained models that are continuously improving their performance and applicability across pre-trained models encompassing multiple languages of North-East India would definitely improve the NLP in North-East India. By doing this, it builds a strong base of these languages based on the similarities in their structures and characteristics for other NLP uses. Thus, by populating pre-training data with various Tibeto-Burman languages, these models can acquire general features for the family, which would enable better transfer learning. This is particularly beneficial for the very low-resource languages which may, at times, have inadequate data for training of individual models. Such multilingual models will generally enhance performance in tasks such a machine translation, text classification, and named entity recognition in multiple languages. Further, this approach can capture cross-lingual features and insights thereby enhancing the study of the linguistic context in North-East India. These models can be further fine-tuned or enlarged if newer data becomes available which will result in continuously improving their performance and applicability across the diverse languages of the region [34].

### 7.2 Speech Technology:

Since most of the languages in North-East India are spoken predominantly in informal forms of communication, it makes the development of a speech recognition and synthesis system even more crucial. These tools could revolutionize language preservation, learning, and interaction in people’s everyday lives. Free text-to-speech conversion systems would make it possible to develop extensive corpora of spoken languages which would be a valuable resource for documentation of endangered languages and additional advancements in the field of NLP. They could also help in designing the voice-based and other interfaces to make digital technologies usable by the illiterate speakers [16]. On the other hand, there is a potential to use speech synthesis to help preserve endangered languages by offering pronunciation help, for instance. These technologies could improve the delivery of information in the healthcare facilities and government services where language may act as a barrier to receiving such information. In addition, use of speech to text and text to speech tools may also help in development of audio books and educational information to aid with literacy. By embarking on speech technologies, one can devise tools that are closer to the linguistic prerequisites and necessities of North-East Indian peoples and may even surpass text- centred approaches in certain applications [35].

### 7.3 Language Documentation:

Using NLP approach in language documentation activities can go a long way in documenting the low-resourced and endangered languages in North-East India. Using such technologies, language professionals and researchers can easily gather, transcribe, and analyse large quantities of spoken and written language. Besides helping in constructing adequate linguistic documentation it also facilitates in developing useful assets for NLP [36]. Better documentation leads to the creation of language models, corpora, and lexicons pertaining to these languages. Furthermore, the use of NLP for documenting content can help in the creation of instructional resources that can be used in raising awareness on language learning and language preservation. Lastly, using NLP in language documentation also benefits the conservation of language heritage as well as improvement of NLP technologies [37].

## 8. CONCLUSION

Developing NLP tools for low resource languages of North-East India poses great challenges as well as great opportunity. Due to the tonal property of the language and the complexity of the morphological and syntactic structures of the Tibeto-Burman languages, new techniques in NLP are required. Emerging techniques like transfer learning, multilingual pre-training, and few-shot learning seem to address the challenge of data deficiency. Integrating the speech technologies might bring significant changes to education, healthcare services, and digital accessibility as most of these languages are mainly oral in nature. The future of these efforts relies on interdisciplinary collaboration, which include the integration of linguistic, anthropological, and computer science knowledge. The effectiveness of this approach is geared towards the creation of technically proficient tools that are also culturally acceptable and sensitive the community needs. Moving forward, it is necessary to examine the ethical concerns to make sure these NLP technologies empower and do not marginalize these linguistic communities. Efforts must continue to strive for the preservation of these languages and ensure not to inadvertently promote language shift. The findings made and methods derived in this context can be useful for low-resource language processing globally and thereby make technology more inclusive. Thus, despite its difficulties, this research area of study is significant not only for advancement of NLP but also for preserving linguistic diversity, rendering it a crucial research area.

## REFERENCES

1. Moral, D. (1997). North-East India as a linguistic area. *Monkhmer Studies*, 43-54.
2. Talukdar, S., & Sarma, S. K. (2024). Enabling Natural Language Processing and AI Research in Low-Resource Languages: Development and Description of an Assamese UPoS Tagged Dataset. *Journal of Electrical Systems*, *20*(3s), 1312-1320.
3. Bose, A., & Majumder, G. (2024). A Case Study on Tools and Techniques of Machine Translation of Indian Low Resource Languages. In *Empowering*

*Low-Resource Languages With NLP Solutions* (pp. 51-85). IGI Global.

4. Evans, N., & Levinson, S. C. (2009). The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and brain sciences*, *32*(5), 429-448.
5. Post, M. W., & Burling, R. (2017). The Tibeto-Burman languages of Northeast India. *The Sino-Tibetan Languages*, *213*, 242.
6. Chhangte, L. (1993). *Mizo syntax*. PhD thesis, University of Oregon.
7. Singh, T. D., & Bandyopadhyay, S. (2010). Statistical machine translation of english–manipuri using morpho-syntactic and semantic information. In *Proceedings of the Association for Machine Translation in the Americas (AMTA 2010).*
8. DeLancey, S. (2015). Morphological evidence for a central branch of Trans-Himalayan (Sino-Tibetan). *Cahiers de linguistique Asie orientale*, *44*(2), 122-149.
9. Wikipedia Contributors. (2024). Meitei script. https://simple.wikipedia.org/wiki/ Meitei_script, Accessed: 2024-07-24.
10. Wikipedia Contributors. (2024). Mizo script. https://en.wikipedia.org/wiki/Mizo_ language, Accessed: 2024-07-24.
11. Wikipedia Contributors. (2024). Ao script. https://en.wikipedia.org/wiki/Ao_language, Accessed: 2024-07-24.
12. Khenglawt, V., Laskar, S. R., Pakray, P., & Khan, A. K. (2024). Addressing data scarcity issue for English–Mizo neural machine translation using data augmentation and language model. *Journal of Intelligent & Fuzzy Systems*, (Preprint), 1-11.
13. Kuiken, F. (2022). Linguistic complexity in second language acquisition. *Linguistics van- guard*.
14. Ashwin, R., Maulik, D., & Maulik, T. (2019). A comparative study of various techniques and challenges for hand written document processing of indian script.
15. Chakma, K., & Das, A. (2021). An overview of parts of speech tagging approaches for indian languages. *ACM Transactions on Asian and Low-Resource Language Information Processing*, *20*(2), 1–33.
16. Susma, T., Pradip, K., Devi, D. (2023). Speech dataset development for a low- resource tibeto-burman tonal language. In *Proceedings of the International Conference on Oriental COCOSDA*, 1–6.
17. Pratapa, A., Bhat, G., Choudhury, M., Sitaram, S., Dandapat, S., & Bali, K. (2018, July). Language modeling for code-mixing: The role of linguistic theory based synthetic data. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1543-1553).
18. Sarkar, S., Das, S., Nath, B., & Mukhopadhyay, S. (2024). A Multilingual Neural Machine Translation Model for Low Resource North Eastern Languages.
19. Kakwani, D., Kunchukuttan, A., Golla, S., Gokul, N. C., Bhattacharyya, A., Khapra, M. M., & Kumar, P. (2020, November). IndicNLPSuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020* (pp. 4948-4961).
20. Khanuja, S., Bansal, D., Mehtani, S., Khosla, S., Dey, A., Gopalan, B., ... & Talukdar, P. (2021). Muril: Multilingual representations for indian languages. *arXiv preprint arXiv:2103.10730*.
21. Xia, M., Kong, X., Anastasopoulos, A., & Neubig, G. (2019). Generalized data augmentation for low-resource translation. *arXiv preprint arXiv:1906.03785*.
22. Lalrempuii, C., & Soni, B. (2023, November). Investigation of Data Augmentation Techniques for Assamese-English Language Pair Machine Translation. In *2023 18th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP)* (pp. 1-6). IEEE.
23. Ziyaden, A., Yelenov, A., Hajiyev, F., Rustamov, S., & Pak, A. (2024). Text data augmentation and pre-trained Language Model for enhancing text classification of low-resource languages. *PeerJ Computer Science*, *10*, e1974.
24. Kumar, R., Takhellambam, M., Lahiri, B., Gope, A., Ratan, S., Mathur, N., & Singh, S. (2023). Collecting Speech Data for Endangered and Under-resourced Indian Languages. In *Proc. 2nd Annual Meeting of the ELRA/ISCA SIG on Under-resourced Languages (SIGUL 2023)* (pp. 14-18).
25. Shanmugasundaram, S. (2022). Gamified crowdsourcing for idiom corpora construction. *Natural Language Engineering*.
26. Bender, E. M., & Koller, A. (2020, July). Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 5185-5198).
27. Yin, W. (2020). Meta-learning for few-shot natural language processing: A survey. *arXiv preprint arXiv:2007.09604*.
28. Lample, G. (2019). Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.
29. Singh, S. M., & Singh, T. D. (2022). Low resource machine translation of english–manipuri: A semi-supervised approach. *Expert Systems with Applications*, *209*, 118187.
30. Tzudir, M., Baghel, S., Sarmah, P., & Prasanna, S. R. (2022). Under-resourced dialect identification in Ao using source information. *The Journal of the Acoustical Society of America*, *152*(3), 1755-1766.
31. Narzary, S., Brahma, A., Nandi, S., & Som, B. (2024). Deep Learning based Named Entity Recognition for the Bodo Language. *Procedia Computer Science*, *235*, 2405-2421.
32. Bird, S. (2020). Decolonising speech and language technology. In *28th International Conference on Computational Linguistics, COLING 2020* (pp.

3504-3519). Association for Computational Linguistics (ACL).

33. Tonja, A. L., Balouchzahi, F., Butt, S., Kolesnikova, O., Ceballos, H., Gelbukh, A., & Solorio, T. (2024). NLP Progress in Indigenous Latin American Languages. *arXiv preprint arXiv:2404.05365*.
34. Lalrempuii, C., & Badal, S. (2023). *Low-Resource Indic Languages Translation Using Multilingual Approaches*, 371–380.
35. Lekshmi, J., Akhila, N., & Vrinda, S. K. (2022). Indian text to speech systems: A short survey, 1–8.
36. Zariquiey, R., Oncevay, A., & Vera, J. (2022, May). CLD² Language Documentation Meets Natural Language Processing for Revitalising Endangered Languages. In *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages* (pp. 20-30).
37. Flavelle, D., & Lachler, J. (2023, May). Strengthening Relationships Between Indigenous Communities, Documentary Linguists, and Computational Linguists in the Era of NLP-Assisted Language Revitalization. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)* (pp. 25-34).