International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 23(s), 2025
https://theaspd.com/index.php

Processing Low-Resource Languages: A Review Of Challenges And Strategies For Inclusive NLP And Sustainable Environment

Dr. Diganta Baishya[1*], Dr. Rupam Baruah[2], Dr. Mousoomi Bora[3], Biswajit Sarma[4]

[1*]*Jorhat Engineering College, dbaishya@gmail.com*
[2]*Jorhat Engineering College, rupam.baruah.jec@gmail.com*
[3]*The Assam Kaziranga University, mousoomi@gmail.com*
[4]*Jorhat Engineering College, eduneristbiswa@gmail.com*

Abstract

The recent advances of Natural Language Processing (NLP) have significantly benefited people across the globe who speak and write some specific languages, commonly termed high-resource languages (HRLs). These are languages with abundant digital resources. One of the most significant areas where NLP has recently made its mark is environment preservation. However, a significant digital divide is observed for low-resource languages (LRLs), which are not rich in terms of digital resources. Even though many LRLs, like Assamese, are very important for the cultural and linguistic identity of many indigenous communities, they remain digitally underrepresented due to a lack of annotated corpora and computational tools. This gap is even more crucial in the environmental domain, where knowledge is very limited even for high-resource languages. Multilingual information is very important for NLP applications related to climate change, disaster management and other issues related to environmental science. This paper reports a study of the challenges and strategies to address the gap in NLP, both in the context of the global and Indian landscape. The paper also highlights the key problems while addressing issues related to environmental science due to the lack of digital resources. Some of the significant factors include the absence of large corpora, labelled datasets and socio-economic factors. The paper proposes to emphasize data collection and digitization, among many other measures, to address this gap. The integration of multilingual pretrained models and transfer learning approaches also provides a pathway for enhancing performance with limited resources. The paper also presents a detailed analysis of resources available for global and Indian languages in addition to proposing a set of strategic actions, including government policies, and others, for inclusive development in NLP and sustainable environment. Bridging the divide and enabling linguistic equity will ensure participation of all sections of the society in the advancement of inclusive technological growth and solutions to the emerging problems.

1. INTRODUCTION

Natural language has long distinguished humans from other species, and with technological progress, communication has undergone significant transformation. The automation of natural language processing (NLP) has simplified many tasks and enabled machines to analyze, understand, and generate languages spoken by humans.

Hirschberg and Manning [1] define NLP as the set of methods that facilitate automatic processing of languages. Chowdhary [2] describes NLP as a discipline concerned with understanding and generating languages spoken by humans. Modern NLP applications have benefited a lot from high-quality research in areas such as emotion detection, speech recognition, and machine translation. Availability of large linguistic datasets, and improvements in machine learning techniques have enabled a huge transformation in the field of NLP [1]. Research into NLP has evolved with the evolution of computational models, neural networks and availability of data. Early research was based on rule-based models and stochastic methods. These rules were primarily handwritten, very complex and required experts to define them. NLP research was primarily limited to tasks like POS tagging and named entity recognition during this stage. However, tasks that require deep knowledge and study were very hard to be implemented. Concepts of support vector machines (SVMs) and decision trees proved to be very useful machine learning tools for classifications and information extraction, but such methods struggled to handle ambiguity and contexts in text. The introduction of deep learning and ability to learn from data revolutionized NLP. The Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) architectures and their variants have been the key factors behind some innovative solutions that have revolutionized the area of NLP. On the other hand, Convolutional Neural Networks (CNNs) have also proven

to be very useful for text classification. However, these models suffer from long-term dependency in text. The Transformer architecture and attention mechanism, addressed long-range dependencies effectively, laying the foundation for models such as BERT and GPT [3] [4] [5] [6]. Social media, emails and huge digital repositories have resulted in an abundance of data required for deep neural networks. This has enabled the models to perform with high accuracy in the field of healthcare, education, and communication. Some of the areas that have achieved benefits include text summarization, named entity recognition, machine translation, emotion detection, text to speech conversion and speech to text conversion. However, NLP is a challenging task due to grammatical and linguistic complexity of natural languages, both in the case of speech and text processing.

NLP has also been applied in the domain of environmental preservation. For example, CLIMATEBERT is a transformer-based language model proposed in this domain, which is further pretrained on over 2 million paragraphs of climate-related texts [7]. NLP research has facilitated some real-world applications in climate governance, including sentiment analysis of climate-related social media discussions and predictive modeling of policy impact [8] [9]. But the technologies available for NLP at present, are primarily focused only on the high-resource languages. The technologies used so far including modern methodologies, make use of a huge amount of data, which is not available for low-resource languages. This imbalance restricts the global inclusivity of language technologies. Addressing these limitations is therefore essential to ensure that the benefits of NLP can be extended equitably across all languages and societies. There are over 7,100 living languages worldwide [10], yet only a fraction have adequate computational resources. While English, Mandarin, and a few European languages dominate existing NLP tools, the vast majority of languages remain underrepresented**.** Many of these languages are spoken by small communities, often in regions with rich cultural and linguistic heritage. This imbalance poses a serious risk of language extinction**,** where languages without technological support gradually lose relevance in digital communication. Addressing this disparity is essential to ensure linguistic inclusivity and the equitable benefits of NLP technologies across all global communities.

This divide is a serious concern for research in the environmental domain, because documented knowledge plays a vital role in critical applications related to climate change, earthquake, pandemic and disaster management. Low-resource languages can provide significant inputs in these areas with localized context and indigenous knowledge. But due to lack of digital resources and lack of NLP applications for LRLs, they have not been able to contribute much in this domain. Addressing this gap therefore, is not only a technological necessity but also a socio-cultural and ethical imperative.

2. LANGUAGE DIVERSITY ACROSS GLOBAL LANGUAGES

Despite the presence of thousands of languages worldwide, NLP applications remain concentrated in just a few. Table 2.1 highlights the top languages by percentage of global speakers:

Table 2.1: Most Widely Spoken Languages Globally [11]

Sl. No.	Language	% of Global Speakers
1	Mandarin Chinese	12.3%
2	Spanish	6.0%
3	English	5.1%
4	Arabic	5.1%
5	Hindi	3.5%
6	Bengali	3.3%
7	Portuguese	3.0%
8	Russian	2.1%
9	Japanese	1.7%
10	Western Punjabi	1.3%

Yet, most NLP tools are designed primarily for English, Chinese, Arabic, Spanish, and French. This mismatch between widely spoken languages and those supported by NLP leaves large populations digitally underrepresented. Magueresse et al. [12] observe that most NLP research is conducted in only about 20 out of approximately 7,126 languages in the world. Programs such as REFLEX-LCTL [13] and the Endangered

Languages Documentation Project [14] confirm that most resources belong to a limited set of languages. This creates a clear divide between high-resource and low-resource languages, where the scarcity of digitized corpora significantly hinders technological development.

Globally, this divide persists. Although languages such as Quechua (Peru), Wolof (Senegal), and Javanese (Indonesia) have millions of speakers, they lack even basic NLP tools such as tokenizers, part-of-speech taggers, or translation systems. Joshi et al. (2020) [15] noted that NLP research was focused on fewer than 20 languages. The primary reason behind this imbalance is the lack of resources and funding for research. Similarly, Blasi et al. (2022) [16] show that training datasets for large language models (LLMs) are mainly available in English. This imbalance is one of the key reasons why native speakers of low-resource languages have not benefited from advanced AI tools. This has a serious consequence to the society at large. Bird (2020) [17] argues that technological support for languages is very essential for preserving the cultural and historical identity of Indigenous communities in Australia and Papua New Guinea. Without availability of digital tools, many of these languages are facing extinction and denial of benefits. However, initiatives such as Masakhane [18] in Africa and AmericasNLP [19] for Indigenous American languages have been very promising.

Importantly, the knowledge about human behavior is incomplete without much information about majority of the languages spoken in the universe. Many important aspects for understanding the environment are excluded, as much of this knowledge is closely linked to LRLs. Some of these areas include indigenous farming methods, water conservation across the globe, techniques of biodiversity preservation across communities, and disaster adaptation strategies. With information collected from each and every corner of the globe the effective NLP tools can be designed for LRLs including tools related to environment safety. Indigenous languages are hub ecological knowledge that helps support biodiversity and sustainable farming. Extinction of these languages or no knowledge about them is a serious concern for developing sustainable solutions to addressing climate change and ecological challenges. For example, Quechua in Peru encodes centuries of agricultural practices in high-altitude ecosystems. On the other hand, there are many Indigenous African languages that preserve oral traditions of soil management and drought prediction. Also, Indigenous American languages often embed knowledge on ecology and medicinal plants. But such important knowledge is not accessible to scientific communities because of lack of digitized corpora [20].

The lack of available resources for low-resource languages limits the participation of the majority of people, their cultural history, and knowledge in this era of technological development. This exclusion not only limits technological development but also prevents coming up with better solutions for global issues like disaster management, early-warning alerts, and environmental policy-making with the help of an inclusive knowledge base. Current NLP tools overlook critical voices by excluding LRLs. Therefore, enriching NLP research on LRLs is not only a linguistic necessity but it is also a crucial step towards domains critical to human survival.

3. LANGUAGE DIVERSITY IN INDIA

India presents a striking example of this global divide. While the country is home to hundreds of languages, NLP resources remain scarce for most of them. Table 3.1 shows the percentage of the population speaking major Indian languages:

Table 3.1: Percentage of Speakers of Major Indian Languages [21]

Language	percentage of Indian Population
Hindi	40.10%
Bengali	8.85%
Marathi	8.18%
Telugu	7.77%
Tamil	6.36%
Gujarati	4.99%
Urdu	5.18%
Kannada	4.84%
Odia	3.51%
Malayalam	2.93%

Punjabi	2.97%
Assamese	1.94%
Maithili	1.12%
English	10.67%

Despite millions of speakers, the majority of Indian languages remain "digital dark zones," with limited or no NLP infrastructure [22]. Existing resources are often small, fragmented, and insufficient for building robust tools [23]. Even widely spoken languages such as Bengali, Tamil, and Marathi lag behind English and Hindi in terms of digital content and NLP tool availability. Studies show that models underperform significantly for most Indic languages due to lack of high-quality corpora [23] [24][25].

This underrepresentation has direct social and economic consequences. About 68% of the internet users in India prefer digital content to be more credible in the local language [26]. Nine out of ten new internet users in India are likely to be Indian language speakers over the next 5 years. [27]. Yet digital services, e-commerce platforms, and educational resources overwhelmingly favor English and Hindi, excluding many speakers of regional languages.

This has created a linguistic digital divide within India, where lack of NLP tools limits participation of all sections of the society. Government initiatives have attempted to address this challenge. For instance, DIKSHA and the National Digital Library of India provide multilingual educational resources, though they typically cover only a subset of constitutionally recognized languages [28]. More recently, Bhashini (launched in August 2022) aims to create scalable translation and speech tools for Indian languages, but its coverage remains limited and heavily dependent on pre-existing corpora [29]. As Dongare [30] notes, standardized processes for data creation are still lacking, preventing systematic progress.

The implications also extend to the environmental domain. Forest conservation traditions in tribal areas, hill protection policies in hilly areas, local disaster-response strategies in flood- and cyclone-prone areas and other innovative and proven strategies have been followed over centuries in India. But the lack of documentation has prevented their incorporation into the strategies involved in disaster management systems. In sum, India mirrors the global trend: despite linguistic richness and millions of speakers, most languages remain digitally marginalized. The challenges faced by Indian languages—resource scarcity, lack of benchmarks, and unequal digital inclusion—are emblematic of the broader struggles of low-resource languages worldwide. Addressing these issues in India will therefore contribute not only to national development but also to the global effort of preserving the environment.

4. CHALLENGES IN DEVELOPING NLP FOR LOW-RESOURCE LANGUAGES

Low-resource languages (LRLs) face persistent and multifaceted challenges that hinder their integration into modern NLP systems. A foundational barrier is the scarcity of digitized text and speech corpora, which are essential for training robust models [15] [31] [32] [33]. In addition, many LRLs are excluded from standard NLP benchmarks, limiting opportunities for evaluation and comparative analysis [32] [34]. Infrastructural gaps such as missing Unicode support, lack of standardized fonts and keyboard layouts further extend the problem [31]. At a broader level, economic and geopolitical incentives are focused on a handful of dominant languages, leaving the majority of the world's linguistic diversity underrepresented [15]. This systematic marginalization denies speakers of LRLs equitable access to digital tools, information, and essential services.

Despite recent advances, several technical and ethical challenges remain. Pretrained models continue to underperform on low-resource languages, often replicating social, cultural, and linguistic biases. Large language models (LLMs) such as GPT-4 are prone to hallucination, producing fluent but factually incorrect outputs. Moreover, nuanced contextual and semantic understanding—particularly of idiomatic or culturally specific expressions—remains limited. As highlighted in recent literature, future research must focus on multilingual and cross-lingual transfer systems, greater model interpretability, and participatory approaches to ensure inclusivity and reliability.

A closer look at the challenges reveals multiple interrelated dimensions:

• Less Investment**:** Large-scale financial and institutional support prioritizes high-resource languages. Magueresse et al. (2020) note that most of the world's 7,000+ languages receive little to no NLP attention, stalling resource creation and model development [12] [35].

• Low Digitization: Limited digital infrastructure restricts corpus generation. Hedderich et al. show that state-of-the-art NLP struggles in such settings due to the absence of large digital datasets [36] [37].
• Lack of Digitized Environmental Corpora: Most of the reports related to climate and ecological analysis are published in English or other high-resource languages. Thus, the local languages lack enough digital content for quality research.
• Absence of Standardized Terminology: Environmental concepts often lack equivalent terms in LRLs. This leads to inconsistent translations that make it difficult to develop a reliable NLP solution for a sustainable environment.
• Illiteracy and Lack of Standardization: High illiteracy rates and non-standardized writing systems complicate text-based resource development. Also, resources for languages spoken by communities with low literacy rate are usually not available [12].
• Demographic and Dialectal Diversity**:** Multiple dialects, scripts, and variants within LRL communities increase modeling complexity [38].
• Lack of Annotated Corpora**:** The absence of labelled datasets for tasks like POS tagging, NER, or parsing remains a core bottleneck [12].
• Benchmark Exclusion: Frequent omission from evaluation frameworks like GLUE and XTREME limits visibility and progress [36].
• Inadequate NLP tools**:** Tokenizers, morphological analyzers, and embeddings tailored for LRLs remain scarce, reducing model performance even when multilingual models such as mBERT or XLM-R are applied [39].
• Cultural and Semantic Nuance Loss**:** Generic models often misinterpret culturally specific expressions, eroding trust [39].
• LLM Limitations: Even cutting-edge LLMs fail to represent LRLs adequately due to training biases and lack of community participation [40].

Table 4.1 : Summary of challenges in low-resource languages

Challenge	Core Issue	Example Case	Reference
Less Investment	Minimal funding for datasets/tools	Indian languages under-resourced	[12]
Low Digitization	Limited internet/devices, few digital texts	Hausa, Igbo corpora scarcity	[36]
Lack of Digitized Environmental Corpora	Limited digital texts	Environmental Corpora	[15] [17]
Illiteracy / Non-standard script	No written form or low literacy → no text data	Endangered oral-only languages	[12]
Dialect / Demographic Diversity	Multiple variants complicate modeling	Creoles in West Africa	[38]
Lack of Annotated Corpora	No labelled datasets for NLP tasks	POS-tagging gaps	[12]
Benchmark Exclusion	No shared evaluation frameworks	Many Indian regional languages	[36]
Tooling Gaps	Missing tokenizers/embeddings	Morphological tools for Punjabi absent	[39]
Semantic/Cultural Nuance Loss	Models misinterpret idioms/metaphors	Translating local metaphors incorrectly	[39]
LLM Underperformance	Biased toward high-resource languages	Burmese or Swahili poorly supported	[40]

In sum, the challenges faced by LRLs span social, infrastructural, and technical dimensions. Addressing them will require strategic investment**,** community-led data collection**,** benchmarks tailored to diverse dialects**, and**

cultural sensitivity in model design. Only through such inclusive and participatory approaches can NLP systems equitably serve the world's full linguistic diversity.

5. CASE STUDY: RESOURCE ANALYSIS FOR GLOBAL AND INDIAN LANGUAGES

High-resource languages such as English and Mandarin Chinese, with very large speaker populations, dominate the digital landscape. These languages benefit from millions of Wikipedia articles, frequent inclusion in standard NLP benchmarks like GLUE, XTREME, and XGLUE, and extensive availability of pretrained models such as BERT, RoBERTa, GPT, and ERNIE [32] [41]. Their strong representation is the result of sustained research attention and infrastructural support over several decades.

In contrast, low-resource languages such as Assamese, Sora, and Khoekhoe remain largely excluded from this ecosystem. They lack digital representation, have few or no Wikipedia articles, are absent from most benchmarking datasets, and rarely have dedicated pretrained models [42] [43]. Even mid-resource languages such as Hindi and Bengali—despite having hundreds of millions of speakers—are underrepresented in global NLP datasets and tools. Nevertheless, recent initiatives like IndicGLUE, IndicNLP, and models such as IndicBERT and MuRIL represent important steps toward bridging this gap [44].

This uneven distribution of resources not only limits technological access for speakers of the majority of the languages but also reinforces existing digital inequalities. Assamese, with a speaker base of over 2 million, remains critically under-resourced, while languages such as Welsh with far fewer speakers benefit from stronger representation due to availability of resources and academic efforts. These findings underscore the urgent need for community-driven, policy-supported, and open-source initiatives that democratize language technologies and safeguard linguistic diversity in the digital era.

The technologies are concentrated more towards high-resource languages. This imbalance carries the risk of denial of service to the millions of people speaking low-resource languages, specially in India and other regions which are houses of thousands of communities. Collecting foundational datasets and fostering open-source collaborations are some of the strategies that are essential for an inclusive growth. The following section highlights some strategic actions to bridge these gaps and to develop NLP applications for low-resource languages.

Moreover, these gaps are even more critical in the domain of environment preservation. LRLs such as Assamese, Santali, and Sora, encode rich ecological knowledge systems. Such knowledge systems include knowledge of agriculture, water management, flood control, earthquake preparedness and biodiversity. But such knowledge remains inaccessible to modern NLP pipelines due to poor digital representation in comparison to high-resource languages like English. Bridging this gap therefore should be among the top priorities of human civilization.

6.SOLUTIONS AND STRATEGIC ACTIONS

Developing NLP for LRLs requires some collaborative efforts. Table 6.1 below highlights a structured framework for addressing the problems related to developing NLP for low-resource languages (LRLs). It must begin with foundational efforts such as data collection and digitization [31] [15]. It is also important to note that, many LRLs are rich repositories of ecological text, environmental knowledge, folk songs, oral traditions, and agricultural records. These repositories are enriched with linguistic resources that often encode centuries of sustainable practices related to agriculture, forest use, and biodiversity preservation. Government policies should be framed to ensure their preservation through digitization and computational modeling. Community engagement is a must for such strategies. This ensures that native speakers participate in the validation and enrichment of linguistic resources and other resources. Such initiatives also help them to maintain their cultural integrity [17] and contribute towards developing inclusive strategies for environment preservation.

For many LRLs, particularly those without standardized scripts, developing or refining orthographies is crucial for text processing and further computational applications. The creation of parallel corpora and the adoption of multilingual pretrained models (e.g., mBERT, XLM-R) can enable transfer learning and enhance performance even with limited data [34]. Furthermore, speech and ASR datasets are vital for oral languages, while crowdsourced data collection platforms provide scalable, community-driven development approaches [45] [46].

Equally important is the establishment of evaluation benchmarks tailored to LRLs, which allows researchers to track progress systematically [32]. Sustained advancement depends on government support and open-source resource sharing, as both play a central role in mobilizing resources and fostering inclusive innovation ecosystems

[47]. Exploring deep learning [48] and multilingual models [49] are some of the important initiatives already undertaken by researchers. Framing government policies to support research initiatives for low-resource languages will play a vital role in this effort [50]. Research initiatives like investigating the minimal size of training dataset required [51] to perform specific tasks will also improve training efficiency.

Collectively, these strategies form a holistic roadmap for equipping LRLs with modern NLP capabilities. Table 6.1 below highlights some of the strategies to develop technologies for low-resource languages.

Table 6.1: Strategies to Develop Technologies for Low-Resource Languages

Action Step	Description	Justification	References
Data Collection & Digitization	Gather texts, audio, and oral narratives	Foundational step	[32] [15]
Community Participation	Engage native speakers for validation	Ensures sustainability and relevance	[10] [41]
Develop Standard Orthography	Create consistent scripts	Essential for text processing	[43]
Build Parallel Corpora	Translate into high-resource languages	Supports machine translation	[34]
Deep learning for LRLs	Make the deep learning methods work with less training data	Will support languages less data	[48]
Use Multilingual Models	Use mBERT, XLM-R	Enables few-shot/zero-shot learning	[34] [49]
Crowdsourcing Platforms	Use web tools and mobile apps	Scalable and community-driven	[45] [17]
Government & Policy Support	Fund initiatives and infrastructure	Essential for resource mobilization	[51]
Open-Source Resource Sharing	Publish data and models	Accelerates innovation and reuse	[15] [17]

7. ENVIRONMENTAL-SPECIFIC STRATEGIC ACTIONS

Many communities speaking the LRLs preserve their knowledge in the form of oral traditions, folklore, and stories. Designing computational models with such knowledge can be very fruitful for sustainability, disaster management, and biodiversity conservation. A few targeted actions are proposed for addressing environmental and ecological challenges in the context of LRLs

The following strategic actions are recommended:

- Development of documented Corpora: Environmental texts, oral narratives, agricultural history, disaster management strategies, folk songs and stories of LRL communities can be collected and documented in the form of corpora. Such corpora can be a game changer in the area of climate text mining, ecological sentiment analysis, and knowledge extraction related to sustainable practices for environmental preservation.
- NLP during Disaster: Many disaster-prone regions rely on local LRLs and local strategies for emergency communication. Such local strategies, if documented can be very innovative and helpful for developing sustainable solutions. Proposing speech-to-text and real-time translation systems for these LRLs can be very useful for the communities for rapid responses during disasters.
- Cross-Language Transfer: The knowledge relevant to environmental science, available in high-resource languages in documented form can be translated and transferred for use with low-resource languages.
- Government policies and funding should focus on preserving the environmental knowledge available in various forms within indigenous communities. Educating the communities to preserve their resources is very important for an inclusive global growth.

8. CONCLUSION AND FUTURE WORK

The digital divide in language technology is not merely a technical challenge but also a socio-cultural, economic and ecological concern. This study highlights the disparities faced by low-resource languages (LRLs), using global trends and Indian languages such as Assamese as case studies. Despite their cultural richness and significant number of speakers, many LRLs remain digitally underrepresented, limiting their presence in NLP applications and digital ecosystems. Bridging this gap requires a holistic approach that combines foundational resource creation, community engagement, and technological innovation. Our analysis emphasize that data collection and digitization remain the most pressing bottlenecks, while the development of standard orthographies, parallel corpora, and speech datasets provides the basis for scalable NLP systems. Strategic adoption of multilingual pretrained models such as mBERT, XLM-R, and IndicBERT demonstrates how transfer learning can mitigate resource scarcity, offering a pathway to inclusive development even with limited training data. Furthermore, evaluation benchmarks tailored to LRLs, along with open-source sharing of resources and tools, ensure transparency and foster innovation within a collaborative ecosystem. Involvement of speakers from across communities is very important for developing NLP tools for all languages. Government policies should be framed to ensure inclusive growth for all languages. Involving LRLs in developing NLP will ensure linguistic equality, preserve cultural identity of the communities across the globe in addition to safeguarding the environment. With involvement of all, we can move towards a more linguistically diverse digital future and green environment.

This study outlines some of the strategic solutions for developing effective NLP research for LRLs, that in turn can facilitate better policies for environment preservation. However, there is a great scope for exploring the area for research. One of the most important point is to design a benchmark dataset for LRLs. In addition to this, there is a good scope of working on cross lingual transfer learning so that knowledge of one high-resource language can be utilized for other languages. It is important that variations across the languages, dialect, cultural context are taken into consideration while designing future NLP applications. Designing scalable frameworks for mobile applications that offers inclusivity of languages and communities is a promising area of research. Only through coordinated, long-term collaborative efforts of institutions, industries and government, can LRLs achieve genuine digital visibility, which can in turn become a part of inclusive strategies for a sustainable environment.

REFERENCES:

[1] J. Hirschberg and C. D. Manning, “Advances in natural language processing,” Science, vol. 349, no. 6245, pp. 261–266, 2015.
[2] K. Chowdhary, Fundamentals of Artificial Intelligence. Springer, 2020.
[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, et al., “Attention is all you need,” in Advances in Neural Information Processing Systems (NeurIPS), vol. 30, 2017.
[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” arXiv preprint arXiv:1810.04805, 2018.
[5] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, “Improving language understanding by generative pre-training,” OpenAI, 2018.
[6] A. Radford et al., “Language models are few-shot learners,” arXiv preprint arXiv:2005.14165, 2020.
[7] ClimateBert: A Pretrained Language Model for Climate-Related Text, Nicolas Webersinke, Mathias Kraus, Julia Anna Bingler, Markus Leippold [https://arxiv.org/abs/2306.09737]
[8] Jessica Obianuju Ojadi , Ekene Cynthia Onukwulu , Chinekwu Somtochukwu Odionu , Olumide Akindele Owulade "Natural Language Processing for Climate Change Policy Analysis and Public Sentiment Prediction: A Data-Driven Approach to Sustainable Decision-Making" Iconic Research And Engineering Journals, 7(3)
[9] Ghiloufi, H., Merveille, N., Mellouli, S. (2025). Exploring Emerging NLP and Machine Learning Methods in Climate Change Discourse Analysis on Social Media: A Systematic Literature Review. In: Cheriet, M., Boucher, JF., Gondim de Almeida Guimarães, L., Frayret, JM. (eds) Accelerating the Socio-Ecological Transition. Springer, Cham. https://doi.org/10.1007/978-3-031-82896-6_7
[10] Eberhard, David M., Gary F. Simons & Charles D. Fennig. 2023. Ethnologue: Languages of the World. 26th edn. Dallas: SIL International.

[11] List of languages by total number of speakers," Wikipedia, 2024. [Online]. Available: https://en.wikipedia.org/w/index.php?title=List_of_languages_by_total_number_of_speakers&oldid=1224953327.
[12] A. Magueresse, V. Carles, and E. Heetderks, "Low-resource languages: A review of past work and future challenges," arXiv preprint arXiv:2006.07264, 2020.
[13] H. Simpson, C. Cieri, K. Maeda, K. Baker, and B. Onyshkevych, "Human language technology resources for less commonly taught languages: Lessons learned toward creation of basic language resources," Collaboration: Interoperability Between People in the Creation of Language Resources for Less-Resourced Languages, vol. 7, pp. 7–11, 2008.
[14] https://www.nsf.gov/funding/opportunities/dli-del-nsf-dynamic-language-infrastructure-neh-documenting
[15] P. Joshi, S. Santy, A. Budhiraja, K. Bali, and M. Choudhury, "The state and fate of linguistic diversity and inclusion in the NLP world," in Proc. ACL, 2020. [Online]. Available: https://arxiv.org/abs/2004.09095.
[16] Blasi, D., Anastasopoulos, A., & Neubig, G. (2021). Systematic inequalities in language technology performance across the world's languages. arXiv preprint arXiv:2110.06733.
[17] S. Bird, "Decolonising speech and language technology," in Proc. 28th Int. Conf. Computational Linguistics (COLING 2020), 2020. [Online]. Available: https://aclanthology.org/2020.coling-main.512.
[18] Wilhelmina, N., Vukosi, M., Tshinondiwa, M., Timi, F., Taiwo, F., Oluwole, A.S., Shamsuddeen, M., Kabongo, K.S., Salomey, O., Freshia, S. and Andre, N.R., 2020. Participatory research for low-resourced machine translation: A case study in African languages. Findings of the Association for Computational Linguistics: EMNLP 2020, pp.2144-2160.
[19] M. Mager, A. Chaudhary, A. Oncevay, et al., "Findings of the AmericasNLP 2021 shared task on open machine translation for indigenous languages of the Americas," in Proc. AmericasNLP Workshop, 2021. [Online]. Available: https://aclanthology.org/2021.americasnlp-1.4.
[20] Flores, Priscila. Exploring the Role of Ecolinguistics in the Highlands of Peru: A study on the relationship between the Quechua language and local agrobiodiversity. Diss. 2022.
[21] "List of languages by number of native speakers in India," Wikipedia, 2024. [Online]. Available: https://en.wikipedia.org/wiki/List_of_languages_by_number_of_native_speakers_in_India.
[22] S. Sinha and S. S. Agrawal, "Situation and challenges of technologies for indigenous languages of India," in Proc. Language Technologies for All (LT4All), Paris, UNESCO Headquarters, 2019, pp. 5–6.
[23] J. Philip, S. Siripragada, V. P. Namboodiri, and C. V. Jawahar, "Revisiting low resource status of Indian languages in machine translation," in Proc. 3rd ACM India Joint Int. Conf. Data Science & Management of Data (8th ACM IKDD CODS & 26th COMAD), 2021, pp. 178–187.
[24] S. Siripragada, J. Philip, V. P. Namboodiri, and C. V. Jawahar, "A multilingual parallel corpora collection effort for Indian languages," arXiv preprint arXiv:2007.07691, 2020.
[25] Baishya, Diganta, and Rupam Baruah. "Part-of-speech Tagging for Low-resource Languages: Activation Function for Deep Learning Network to Work with Minimal Training Data." ACM Transactions on Asian and Low-Resource Language Information Processing 23, no. 5 (2024): 1-31.
[26] SARMAH, PRIYANKOO. "Indian languages and language technology.
[27] A. Ankush and N. Sawant, "Digital marketing in Indian regional languages: An overview," Int. J. Advance and Applied Research (IJAAR), vol. 9, no. 3, pp. 580–586, 2022.
[28] "Digital infrastructure for knowledge sharing," Wikipedia. [Online]. Available: https://en.wikipedia.org/wiki/Digital_Infrastructure_for_Knowledge_Sharing.
[29] "Bhashini," Wikipedia. [Online]. Available: https://en.wikipedia.org/wiki/Bhashini.
[30] P. Dongare, "Creating corpus of low resource Indian languages for natural language processing: Challenges and opportunities," in Proc. 7th Workshop on Indian Language Data: Resources and Evaluation, 2024, pp. 54–58.
[31] A. Anastasopoulos and G. Neubig, "Pushing the limits of low-resource morphological inflection," arXiv preprint arXiv:1908.05838, 2019.
[32] J. Hu, S. Ruder, A. Siddhant, G. Neubig, O. Firat, and M. Johnson, "XTREME: A multi-lingual benchmark for evaluating cross-lingual generalization," in Proc. ICML, 2020.

[33] Baruah, N., Gogoi, A., Sarma, S.K., Borah, R. (2021). Utilizing Corpus Statistics for Assamese Word Sense Disambiguation. In: Thampi, S.M., Gelenbe, E., Atiquzzaman, M., Chaudhary, V., Li, KC. (eds) Advances in Computing and Network Communications. Lecture Notes in Electrical Engineering, vol 736. Springer, Singapore. https://doi.org/10.1007/978-981-33-6987-0_23
[34] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 8440–8451, Online. Association for Computational Linguistics.
[35] F. Philippy, "NLP for low-resource languages," in European Language Data Space Country Workshop, Luxembourg, Jun. 19, 2024.
[36] M. A. Hedderich, L. Lange, H. Adel, J. Strötgen, and D. Klakow, "A survey on recent approaches for natural language processing in low-resource scenarios," in Proc. NAACL-HLT, 2021, pp. 2545–2568.
[37] Zhong, T., Yang, Z., Liu, Z., Zhang, R., Liu, Y., Sun, H., ... & Liu, T. (2024). Opportunities and challenges of large language models for low-resource languages in humanities research. arXiv preprint arXiv:2412.04497.
[38] A. Joshi, D. Kanojia, H. Lent, H. Kaing, and H. Song, "Connecting ideas in 'lower-resource' scenarios: NLP for national varieties, creoles and other low-resource scenarios," arXiv preprint arXiv:2409.12683, 2024.
[39] P. Pakray, A. Gelbukh, and S. Bandyopadhyay, "Natural language processing applications for low-resource languages," Natural Language Processing, vol. 31, no. 2, pp. 183–197, 2025, doi: 10.1017/nlp.2024.33.
[40] J. N. Pava, C. Meinhardt, H. B. U. Zaman, T. Friedman, S. T. Truong, D. Zhang, V. Marivate, and S. Koyejo, "Mind the (language) gap," 2025.
[41] HuggingFace, "Model Hub," 2025. [Online]. Available: https://huggingface.co/models..
[42] "Ethnologue," Ethnologue. [Online]. Available: https://www.ethnologue.com/
[43] "List of Wikipedias," Meta Wikimedia. [Online]. Available: https://meta.wikimedia.org/wiki/List_of_Wikipedias.
[44] "AI4Bharat," IIT Madras. [Online]. Available: https://ai4bharat.iitm.ac.in/.
[45] R. Sennrich et al., "Neural machine translation of rare words with subword units," in Proc. ACL, 2016.
[46] Mozilla, "Common Voice datasets," 2022. [Online]. Available: https://commonvoice.mozilla.org/datasets..
[47] Lewis, Melvyn & Simons, Gary. (2010). Assessing endangerment: Expanding Fishman's GIDS. Revue Roumaine de Linguistique. 55. 10.1017/CBO9780511783364.003..
[48] Baishya, D., & Baruah, R. (2022, November). Recent Trends in Deep Learning for Natural Language Processing and Scope for Asian Languages. In 2022 International Conference on Augmented Intelligence and Sustainable Systems (ICAISS) (pp. 408-411). IEEE.
[49] Winata, Genta Indra, Guangsen Wang, Caiming Xiong, and Steven Hoi. "Adapt-and-adjust: Overcoming the long-tail problem of multilingual speech recognition." arXiv preprint arXiv:2012.01687 (2020).
[50] M. P. Lewis and G. F. Simons, Sustaining Language Use. Dallas, TX: SIL International, 2017.
[51] Baishya, D., & Baruah, R. (2023, March). Towards Minimal Training for Low Resource Languages. In 2023 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS) (pp. 876-880). IEEE.