Bridging the Gap: Towards Linguistic Resource Development for the Low-Resource Lambani Language

Ashwini Dasare[2(✉)], Amartya Roy Chowdhury[1(✉)], Aditya Srinivas Menon[3(✉)], Konjengbam Anand[1], K. T. Deepak[2(✉)], and S. R. M. Prasanna[1(✉)]

[1] Indian Institute of Technology Dharwad, Dharwad, India
{amartya.chowdhury,konjengbam.anand,prasanna}@iitdh.ac.in
[2] Indian Institute of Information Technology Dharwad, Dharwad, India
{ashwini,deepak}@iiitdwd.ac.in
[3] Indian Institute of Information Technology Kottayam, Kottayam, India
adityasrinivas20bcs8@iiitkottayam.ac.in

Abstract. Language technology development is crucial for many downstream applications such as machine translation and language understanding. The lack of linguistic resources makes it challenging for technology development of under-resource languages. This paper aims at developing linguistic tools for Lambamni, an under-resourced tribal language of India through corpora creation, annotation, and transfer learning from contact language. Based on the annotated corpora, we develop the Lambani language tagset and our investigation focused on various methods for developing a Part-of-Speech (POS) tagger and also creating a morphology dictionary for Lambani. A total of eight BIS tagset is found to be present for Lambani language. The experimental results revealed that the statistical approach with GMM-HMM (Gaussian Mixture Model - Hidden Markov Model) achieved POS tagging accuracy of 96% despite the limited dataset containing 6,893 sentences. This success in a low-resource setting highlights the promising potential of GMM-HMM in overcoming challenges posed by the scarcity of annotated data in under-resourced languages. The experiments not only showcase the effectiveness of the proposed methods for low-resource language processing but also shed light on their applications and open new directions for research in language revitalization and the development of digital tools for zero-resource languages.

Keywords: Langauge technology development · Natural language understanding · Lambani · POS tagger · Morphological analysis

1 Introduction

India is a linguistically diverse nation with over 22 officially recognized regional languages [20] and multiple spoken languages. These languages belong to different language families having unique characteristics, including Indo-Aryan,

A. Karpov et al. (Eds.): SPECOM 2023, LNAI 14339, pp. 127–139, 2023.
https://doi.org/10.1007/978-3-031-48312-7_10

Dravidian, Austroa-Asiatic, Sino-Tibetan, and others [8]. While major Indian languages such as Hindi, Kannada and Tamil have abundant linguistic tools and resources [3,19,22,27], there are many widely spoken low-resource languages that do not have written scripts and linguistic tools such as Lambani, Soliga [6] and Mundari.

Technology plays a vital role in language preservation, offering digital tools like audio and video recording devices, online archives, and language documentation software to record and archive endangered languages for future generations. Language apps and online platforms further aid in language learning and revitalization efforts, providing accessible resources for those interested in studying these languages.

Linguistic Resource (LR) for a language typically encompasses various components that facilitate the development, study, and analysis of that particular language. These resources comprise corpora from diverse sources, lexicons, grammar, phonetics and phonology resources, and morphological analysis tools. Well-established Indian languages like Kannada and Hindi have abundant linguistic resources, such as dictionaries, Part of Speech (POS) taggers, morphological tools, and datasets for Natural Language Processing (NLP) tasks while low-resource languages do not have such facilities.

Globalization, urbanization, cultural assimilation, and limited intergenerational transmission threaten many tribal languages. Endangered tribal languages are more than mere communication tools; they are integral to the identity, worldview, and cultural expression of indigenous communities. Protecting endangered tribal languages is crucial to preserve and revitalize indigenous communities' unique linguistic and cultural heritage worldwide. These languages hold valuable knowledge, history, and traditional practices passed down through generations. Hence, efforts to protect and preserve these languages are essential for the well-being of affected communities and for upholding the diverse richness of human languages and cultures.

Preparing the language corpus for low or zero-resource languages is a challenging and time-consuming task. This is particularly true for languages like Lambani, which lack their own script, making manual tagging a significant hurdle in data annotation, and corpus preparation. In this paper, language preservation activity of Lambani language through technological development is discussed.

The Lambani community, also known as the Banjara community, is culturally rich with a nomadic lifestyle and unique traditions [7,21,28]. They have a fascinating history that spans different regions of India, primarily residing in Karnataka, Andhra Pradesh, Telangana, Maharashtra, and Tamil Nadu. There have been few efforts towards technology building for the Lambani language, such as Machine translation [9], and Text to speech synthesis [10]. But to the best of our knowledge, no literature was found regarding basic linguistic tools for Lambani such as morphological analyzer and POS tagger. This works details the effort to build a POS tagger and a Morphological analyser for Lambani language.

The key contributions of this work are as follows:

- We address the problem of developing linguistic technologies for low-resource languages.
- We create lexical corpora for Lambani language by collecting and translating text from various sources.
- Tagset creation and analysis for Lambani language from the created lexical corpora.
- Development of POS tagger for low-resource languages.
- Development of morphology dictionary from a given text corpora.

The rest of the paper is summarized as follows. A brief overview of earlier works in related area is presented in Sect. 2. The proposed approach for Lambani linguistic technology development is presented in Sect. 3. Section 4 details the evaluation of the developed tools and Sect. 5 concludes the work.

2 Related Works

There have been substantial efforts for the development of linguistic tools of Indian languages for various NLP applications. However, limited linguistic resources, such as dictionaries and part-of-speech taggers, make it difficult to develop high-quality NLP applications for under-resourced languages [29]. The current approaches focus on the development of two broad categories of linguistic tools: POS tagger [5,13,15] and morphological analyzer [4,12].

2.1 POS Tagger

POS tagger development works may be classified into (1) rule-based approaches [2,4,12], (2) statistical approaches [13,15,24], and (3) deep learning-based approaches [11,26]. Antony et al. [5] work on different POS taggers for Indo-Aryan languages like Hindi, Bengali, and Panjab, while Merin et al. [14] discuss various tagging methodologies for Dravidian languages such as Kannada, Telugu, Malayalam, and Tamil languages. Srivastava et al. [26] introduced a Deep Learning (DL)-based unsupervised POS tagging method for Sanskrit, employing character-level n-grams. Deshmukh et al. [11] proposed a deep learning-based POS tagger and a Bi-LSTM-based POS tagger, respectively, for Marathi language. This paper works on developing POS tagger for Lambani languages leveraging these extant techniques.

2.2 Morphological Analyzer

There has been considerable work on Morphological Analysers and generators for Indian Languages. Antony et al. [4] proposed rule based morphological analyzer for Kannada. Veen Dixit et al. [12] developed a rule-based spell checker for Marathi Language. However, data scarcity of under-resource language makes it challenging to develop morphological analyzers as they require diverse data to capture language nuances [29].

2.3 Lambani Lingustic Technology

Due to the lack of script, there has not been much written literature found in the Lambani language. As a result, limited work has been carried out for development of Lambani linguistic tools. To overcome the limitations of data scarcity, researchers [29] propose text corpus creation for under-resource language through the use of a contact language. Amartya et al. [9] worked on developing machine translation methods to translate English text to Lambani for Lambani corpora generation. Ashwini et al. [10] proposed the use of Text To Speech synthesis tools for creating Lamabani dataset. This work extend the above works to generate Lambani corpus through the use of Kannada as a contact language.

3 Proposed Approach

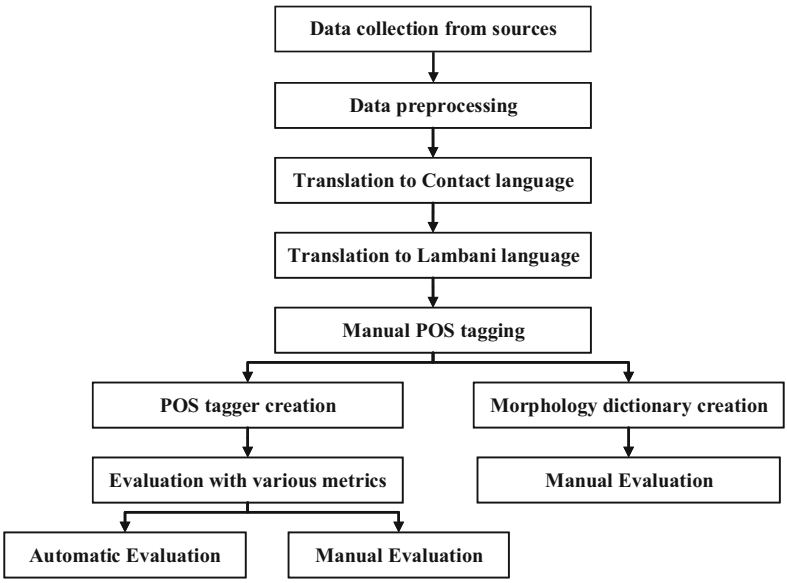

Fig. 1. Architectural overview of the system.

In this section we introduce our proposed system to develop linguistic tools for Lambani. The architectural overview of the system is shown in Fig. 1. The overall process consists of the following steps: (1) Data collection; (2) Data preprocessing; (3) Translation to contact language; (4) Manual POS tagging; (5) POS tagger creation; and (6) Morphology analysis. system undergo the following steps:

3.1 Data Collection

The main objective of this study is to create linguistic resources specifically for Lambani. To overcome the limitation of data scarcity for Lambani language,

this step proposes the creation of Lambani language corpora through transfer learning to use in language tool development. The entire data collection process may be summarised in six steps:

- **Gathering text from various sources**: We utilise the Optical Character Recognition (OCR) feature of Adobe Reader to extract sentences from Lambani-based textbooks [7]. Additionally, we extract English texts from the English subject of the National Council of Educational Research and Training (NCERT) textbooks [1]. Our focus lies specifically on English language textbooks intended for lower and middle schools, encompassing classes I to VI. Further, a linguist manually created 1000 sentences using the Swadesh list [17]. This list comprises a set of basic English words that cover fundamental concepts of English grammar, such as pronouns or verbs.
- **Preprocessing**: The extracted text often contains a significant amount of noise, posing challenges for accurate translation by native Lambani speakers. To address this issue, the extracted texts are further subjected to the following preprocessing methods to obtain a clean corpus.
  - It is observed that native Lambani speakers generally communicate using short simple sentences. So, sentences containing fewer than three words and more than eight words are discarded to avoid lengthy sentences.
  - Incomplete sentences provide noisy information and are removed.
  - Manual checking of the text was carried out by a linguist to remove syntactically or semantically incorrect sentences.
  - Sentences containing symbols, URLs and unknown characters are removed.
- **Relevancy pruning**: The sentences are ranked based on relevancy, where 1 is assigned to relevant sentences and 0 otherwise. For example, sentences containing controversial statements including political statements were marked as irrelevant since they are not used in conversations to carry out daily activities. After the sentences have been ranked the relevant sentences are extracted, and the rest of them are discarded. After this step, around 80% of the sentences are retained out of the total 36,000 sentences.
- **Translation to contact language**: For this study, Lambani speakers from northern Karnataka state are considered and they are fluent in both Kannada and Lambani languages. So, Kannada is chosen as a contact language. The English sentences are translated into Kannada by a bilingual English-Kannada speaker. The translated text is validated by another bilingual Kannada-English speaker.
- **Contact language to Lambani Translation**: The Kannada sentences are manually translated to Lambni by a native Lambani speaker who is familiar with Kannada. The translated sentences are written in the Kannada script.
- **Quality checking and correction**: The translated sentences are manually checked and incorrect ones are rectified.

3.2 Developing Lambani Linguistic Resources (LLR)

The linguistic development efforts primarily revolve around creation of essential resources such as a POS tagger and morphological dictionary. These resources

would greatly assist in the development of computational tools for the Lambani language.

Lambani POS Tagger. POS tagging is a valuable tool in natural language processing (NLP) as it helps algorithms understand the grammatical structure of sentences and disambiguate words with multiple meanings. It is commonly used to determine the lexical categories and convey the semantics of each word in a sentence. For example, let us take a look at the following sentences.
Sentence 1: *I saw a bear in the forest.*
Sentence 2: *Please bear with me during this difficult time.*
In these two sentences, even though the word "bear" is spelled and pronounced the same, its meaning and POS tag differ based on the context. Sentence 1 refers to the animal "bear", where "bear" is a noun. Sentence 2, however, uses "bear" as a verb, indicating the act of enduring or tolerating. Understanding the POS tag of the word "bear" in both of these sentences helps to disambiguate the meaning. Accurate POS tagging is essential to enhance the performance of these language-processing algorithms and enables the development of various language-based applications.

Manual POS Tagging. As Lambani spoken in northern Karnataka uses Kannada script to write, we propose using Kannada POS tagging rules as a foundation to develop Lambani POS tagger. Utilising the expertise of native Lambani speakers proficient in both English and Kannada, we conducted manual annotations for POS tagging using the standards POS tagset developed by the Bureau of Indian Standards (BIS) [18]. The POS knowledge of the created parallel text corpus comprising English, Kannada, and Lambani is used to annotate the Lambani text corpus. The manual annotation and evaluation by native Lambani speakers ensure the reliability and accuracy of the POS tagging model, providing a strong foundation for further linguistic exploration and application. This meticulous annotated corpus serves as a gold standard for subsequent analysis and testing of the POS tagging model. Table 1 shows examples of Lambani POS along with meaning of words in English.

Developing POS Tagger. We compare various methods for POS tagging for developing Lambani POS tagger, including rule-based, Artificial Intelligence (AI) based, Machine Learning (ML) based, and Deep Learning (DL) based approaches. Rule-based methods for POS tagging involve manually creating linguistic rules, but this is time-consuming, error-prone, and requires language experts. An alternative rule-based approach uses a model to learn rules from a training corpus, leading to AI-based methods. Artificial Intelligence methods employ Hidden Markov Models (HMMs) to automate POS tagging, showing good results. However, the trend is shifting towards Machine Learning (ML) approaches like Naive Bayes, SVMs, and CRFs and Deep Learning (DL) based approaches like Long Short-Term Memory (LSTM) networks, Gated Recurrent Units (GRUs), Convolutional Neural Networks (CNNs) and Transformers. Both these approaches aim to learn the patterns and relationships between words and their corresponding POS tags.

Table 1. Lambani tagset along with examples, English translation and transliteration.

Sl no	Category	Label	Lambani words	English translation
1	Noun	NN	ಭೇಸ್ (bhesi), ಬಳದ (balada)	buffalo, ox
2	Pronoun	PRP	ಏಕ್ (ek), ಮೋರ (vora)	one, his
3	Verb	VB	ಖಾರ್ಚ (khaaricha), ಖುಟಗೋಚ (khutagocha)	eating, finished
4	Adjective	JJ	ಮೋಟೋ (moto), ಆಚೊ (aacho)	big, good
5	Adverb	RB	ಅಚಾನಕ್ (achaanak), ಅತ್ತಜ (attaja)	suddenly, here
6	Conjunction	CCD	ನಿಕಾಣಿ (nikaani), ನಿತರ್ (nitar)	or, but
7	Postposition	PSP	ಮಾಯೀರ (maayira), ಉಪರ (upara)	inside, above
8	Particles	RPD	ಮಾಯಿ (Maayi), ಓರ್ (orr)	in, of

HMM. Hidden Markov Model (HMM) is a stochastic technique used for POS tagging that assigns tags to words based on the most frequent tag in the training data. It follows a step-by-step procedure, extracting unique words, calculating tag occurrence counts, and initializing emission and transmission matrices. These matrices represent probabilities of word-tag observations and tag transitions. The Viterbi algorithm is used to find the most probable sequence of POS tags.

RNN (Recurrent Neural Network). The paper aims leverage different configurations of RNN and LSTM to build a POS tagger for Lambani language. The model implementation involves two LSTM layers, each with 128 neurons, and an output layer with Linear and Softmax components.

BERT. Additionally, the paper explores the use of pre-trained embeddings from a fine-tuned BERT model trained on approximately 29K sentences. Pretrained word or sentence embeddings have become essential in Natural Language Processing. Transformer architectures use Masked Language Modeling (MLM) to train the encoder on text corpora, providing embeddings for downstream tasks like POS tagging. However, these models require large training datasets, which can be challenging for low-resource languages like Lambani. To address this, we will explore two approaches: using multilingual transformers trained on diverse data and reducing the number of parameters to lower data requirements.

Creating Lambani Morphological Dictionary. Identifying root words and affixes are crucial to understanding the fundamental meaning and lexical properties of a word. Table 2 shows examples of English, Hindi, and Lambani words along with their respective root words, prefixes, and suffixes.

Table 2. Examples of root forms and affixes of words in English, Hindi and Lambani.

	Word	Prefix	Root	Suffix
English	unhappiness	un-	happy	-ness
Hindi	किताबें		किताब	ें
Lambani	ಕಾಗದೇನ(kaagadena)	-	ಕಾಗದ(kaagada)ೇನ	

The English word “unhappiness,” has the root word is “happy,” while the prefix “un-” and the suffix “-ness” modify its meaning and grammatical function. Similarly, in the Hindi word "किताबें" (books), the root word is "किताब", represents “book” in English. The suffix "ें" indicates plurality, making the word refer to multiple books. Moreover, a Lambani word, "ಕಾಗದೇನ" (pronounced as “kaagadena”). The root word in this case is "ಕಾಗದ" (pronounced as “kaagada”) meaning “paper” in English. Additionally, the suffiix "ೇನ" (pronounced as “een”) modifies the word’s significance.

Building Affix Lexicon. To handle the lexicon specific to the Lambani language, we follow the following steps:

- Vocabulary construction: A vocabulary is constructed that contains all the distinct word forms encountered in the corpus.
- Data cleaning: Non UTF-8 Kannada characters are removed. Additionally, punctuations are also filtered.
- Stemming: As labelled dataset for stemming is not available, the unsupervised Morphessor tool [25] is used for morphological segmentation to get the stem/root words and affixes. The algorithm is based on a set of rules which are applied iteratively until we get the base form of the word. Morphessor uses dynamic programming based Viterbi algorithm to take cleaned vocabulary as input and trains a model that segments words to get stem/root words and affixes.

Table 3 examples of Lambani words along with their POS and morphological affixes obtained after performing morphology analysis.

Table 3. Lambani dictionary after performing morphology analysis.

Lambani word	POS tag	English meaning	English transliteration	Morphology affixes	
ಅಡಗೋರ	NOUN	cooking	adagira	ಅ	ಡಗೋರ
ಅಬೆ	ADVERB	now	abe	ಅ	ಬೆ
ಅತರಾ	DETERMINER	this much	ataraa	ಅತ	ರಾ
ಅಂತೆಮ	ADPOSITION	in the end	antema	ಅಂತ	ೆಮ
ಕಾಳೋ	ADJECTIVE	black	kaalo	ಕಾ	ಳೋ
ಅನಸಾವಚ್	PARTICLE	to feel	anasaavacha	ಅನಸಾವ	ಚ್
ಆರೋಚು	VERB	am coming	aarochu	ಆ	ರೋಚು
ಆದ್ರ	CONJUNCTION	but	aadra	ಆ	ದ್ರ

4 Evaluation

4.1 Dataset Description

The description of the dataset is shown in Table 4. The dataset contains 29,358 sentences collected from various sources of Lambani text. Out of these, 6,893 sentences were manually tagged and divided into training and testing sets using 5-fold cross-validation.

Table 4. Data statistics.

Sl. No.	Total number of sentences
Number of sentences collected	29,358
Number of manually POS tagged sentence	6,893

4.2 Distribution of POS Tags

The distribution of the POS tags is summarised in Table 5. Upon manual labelling of 31640 words, it is inferred that Lambani has 8 part-of-speech tags present, namely Adjective (JJ), Adverb (RB), Conjunction (CCD), Particle (RPD), Noun (NN), Postposition (PSP), Pronoun (PRP) and Verb (VB). It can be observed from Table 5 that we are getting the highest distribution of tags in case of Verb (VB) followed by Noun (NN).

Table 5. Distribution of BIS POS tags in the dataset.

BIS POS Tag	Count (Manual tagging)	Count (GMM-HMM tagging)
Adjective (JJ)	2,743	2,458
Adverb (RB)	1,923	1,727
Conjunction (CCD)	254	296
Particle (RPD)	93	90
Noun (NN)	7,057	7,577
Postposition (PSP)	1,429	1,299
Pronoun (PRP)	6,729	6,496
Verb (VB)	11,412	11,662

4.3 Baseline

For evaluating the performance of POS tagging we use bi-directional RNN based tagger as the baseline. RNN is useful for sequence labelling with variable length inputs. The baseline is compared with BERT based and GMM-HMM based POS tagger. During model training the maximum sequence length is kept at 150 for both the RNN and BERT based models. The training batch size is kept at 32, and

a beam size of 5 is adopted. The baseline model contains only 1 RNN layer with an embedding dimension of 768. In case of BERT based models both the encoder and decoder contain 6 layers. For the feed-forward neural network we have used 1024 inner states. Both the encoder and decoder contain 4 heads in each attention layer block. The attention dropout and the dropout applied in the feed forward network is kept constant at 0.1. Both the RNN and BERT are trained using the Adam optimizer. Other than the straightforward RNN and BERT based models we have also conducted experiments using DistilBERT [23] and MicroBERT [16]. DistilBERT uses the concept of knowledge distillation where a large and complex model (BERT) is used to train a smaller and compact model by transferring its knowledge to the smaller model. Whereas MicroBERT uses multitask learning to reduce the model size. MicroBERT has only 1.29 million parameters, thereby making it a better alternative to BERT. The model configurations to both these models are kept unchanged as their default values.

4.4 Evaluation Metrics

To determine the performance of the proposed automatic POS tagger we adopt accuracy, precision, recall and f1 score as the evaluation metrics. The metrics are defined as follows:

- **Precision** is defined as the ratio of total number of correctly predicted POS tags by total number of predicted tags.
- **Recall** is defined as the ratio of total number of correctly predicted POS by the sum of correctly predicted tags and the number of missed tags.
- **F1-score**: Given precision and recall, F-score is defined as follows:

$$F1 - score = 2 * (Precision * Recall)/(Precision + Recall) \quad (1)$$

- **Accuracy** is defined as the ratio of the total number of correctly predicted POS tags to the total number of tags in the dataset.

4.5 Results

Table 6. Result obtained on various models.

Models	Accuracy	Precision	Recall	F1-score
GMM-HMM	**0.96**	**0.95**	**0.96**	**0.96**
RNN	0.87	0.87	0.87	0.87
BERT+RNN (BRNN)	0.88	0.88	0.88	0.88
Distillbert (D)	0.86	0.86	0.86	0.86
Distillbert + RNN (DRNN)	0.88	0.88	0.88	0.88
Microbert (M)	0.84	0.84	0.84	0.84
Microbert + RNN (MRNN)	0.89	0.89	0.89	0.89

In this section we report the experimental results based on accuracy, precision, Recall and F1-score. Table 6 shows the performance comparison of POS tagging of various methods adopted. The highest metrics compared with the baseline model are highlighted as bold numbers.

POS Taggers Evaluation. We are getting an accuracy of 87% on our baseline model. From Table 6 we can notice that we are getting the highest accuracy of 96% in the case of GMM-HMM which is almost 10% improvement in performance over the baseline model. This may be due to the models ability to handle data sparsity. GMM-HMM tries to learn the joint probability between the words and its corresponding POS tags. Due to its probabilistic approach the model does not assign zero probabilities to unseen word-POS combinations. Moreover, GMM-HMM uses shared parameters across all the states in HMM. This reduces the total number of parameters. In the case of Distillbert (D) we are getting an accuracy of 86% which is an 1% reduction in performance over the baseline model. We are getting the worst performance in the case of Microbert (M). Although M has very few parameters, it is not able to map the POS tags with its corresponding words.

From Table 6 it is quite evident that we are getting a performance improvement when RNN is trained along with BERT models. If we compare between the base BERT models and BERT models that use pre-trained embeddings, we are getting significant improvement while using pre-trained embeddings. As BERT is pre-trained on large amounts of data it was able to capture semantic relationships between various words. Moreover BERT uses contextual embeddings, meaning the embedding of a word depends on the context of the sentence. The BERT+RNN models are almost similar in performance except MRNN which is giving a 1% improvement.

5 Conclusion and Future Work

This paper presents a seminal work to develop a linguistic resource for the under resourced Lambani language. The work involves creating a lexical corpora, a POS tagset, POS tagger, a lexicon dictionary and morphology analyzer for Lambani. We adopt a transfer leaning approach of using parallel corpora in English and Kannada along with Kannada linguistic rules for the work. Upon manual POS tagging of 31640 words, it is observed that the Lambani tagset consists of eight POS tags specified in the BIS tagset. Numerous experiments were conducted to develop an accurate POS tagger that works well with low-resource corpora. For POS tagging, the GMM-HMM approach outperforms the tested methods and gives an accuracy of 96% for POS tagging task. The future efforts will focus on expanding the manually collected parallel corpus in Lambani, both in terms of its size and the amount of annotated POS tags. We will also focus on other variations of BERT like multilingual BERT finetune on the Lambani sentences. The development of a comprehensive Lambani dictionary and further enhancements to the POS tagger will be pursued as well.

Acknowledgement. We would like to thank Prashant Bannulmath, Sunita Rathod, Rajeshwari Naik and Sunil Rathod for helping us in developing the Lambani POS corpus. The authors would also like to thank "Anatganak", high-performance computation (HPC) facility, IIT Dharwad, for enabling us to perform our experiments, and Ministry of Electronics and Information Technology (MeitY), Govt. of India, for supporting us through the "Speech to Speech translation for tribal languages" project.

References

1. National Council of Educational Research and Training. https://ncert.nic.in/textbook.php
2. Aggarwal, N., Randhawa, A.K.: A survey on parts of speech tagging for Indian languages. In: IJCA Proceedings on International Conference on Advancements in Engineering and Technology, ICAET 2015, vol. 3, pp. 29–31 (2015)
3. Anand Kumar, M., Dhanalakshmi, V., Soman, K., Rajendran, S.: A sequence labeling approach to morphological analyzer for Tamil language. Int. J. Comput. Sci. Eng. **2**(06), 1944–1951 (2010)
4. Antony, P., Kumar, M.A., Soman, K.: Paradigm based morphological analyzer for Kannada language using machine learning approach. Int. J. Adv. Comput. Sci. Technol. (2010). ISSN 0973–6107
5. Antony, P., Soman, K.: Parts of speech tagging for Indian languages: a literature survey. Int. J. Comput. Appl. **34**(8), 0975–8887 (2011)
6. Boopathy, S.: Languages of Tamil Nadu: Lambadi, an Indo-Aryan dialect. Census of India 1961, Tamil Nadu ix, part XII (1972)
7. Burman, J.R.: Ethnography of a Denotified Tribe: The Laman Banjara. Mittal Publications (2010)
8. Chandramouli, C., General, R.: Census of India 2011. Provisional Population Totals. Government of India, New Delhi, pp. 409–413 (2011)
9. Chowdhury, A., Deepak, K.T., Prasanna, S. M.: Machine translation for a very low-resource language - layer freezing approach on transfer learning. In: Proceedings of the Fifth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2022), pp. 48–55. Association for Computational Linguistics, Gyeongju (2022)
10. Dasare, A., Deepak, K.T., Prasanna, M., Samudra Vijaya, K.: Text to speech system for lambani - a zero resource, tribal language of India. In: 2022 25th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA), pp. 1–6 (2022)
11. Dhumal Deshmukh, R., Kiwelekar, A.: Deep learning techniques for part of speech tagging by natural language processing. In: 2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA), pp. 76–81 (2020)
12. Dixit, V., Dethe, S., Joshi, R.K.: Design and implementation of a morphology-based spellchecker for Marathi, and Indian language. Arch. Control Sci. **15**(3), 301 (2005)
13. Ekbal, A., Bandyopadhyay, S.: Part of speech tagging in Bengali using support vector machine. In: 2008 International Conference on Information Technology, pp. 106–111 (2008)
14. Francis, M.: A comprehensive survey on parts of speech tagging approaches in Dravidian languages. In: The IIER International Conference, Beijing, China, 26 July 2015 (2015)

15. Gadde, P., Yeleti, M.V.: Improving statistical POS tagging using linguistic feature for Hindi and Telugu. In: ICON, pp. 1–8 (2008)
16. Gessler, L., Zeldes, A.: MicroBERT: effective training of low-resource monolingual BERTs through parameter reduction and multitask learning. In: Proceedings of the The 2nd Workshop on Multi-lingual Representation Learning (MRL), pp. 86–99. Association for Computational Linguistics, Abu Dhabi (Hybrid) (2022)
17. Hymes, D.: Morris swadesh. Word **26**(1), 119–138 (1970)
18. of Indian Standard, B.: Linguistic resources - pos tag set for Indian languages - guidelines for designing tagsets and specification. https://tdil-dc.in/tdildcMain/articles/134692DraftPOSTagstandard.pdf
19. Kumar, D., Singh, M., Shukla, S.: FST based morphological analyzer for Hindi language. Int. J. Comput. Sci. **9** (2012)
20. Metry, K.: tribal languages in 8th schedule. AGPE Royal Gondwana Res. J. Hist. Sci. Econ. Polit. Social Sci. **2**(1), 19–30 (2020)
21. Naik, C., Naik, D.P.: Banjara stastical report Karnatka state, India (2012)
22. Prathibha, R., Padma, M.: Development of morpholoical analyzer for kannada verbs. In: IET, pp. 22–27 (2013)
23. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108 (2019)
24. Sarkar, K., Gayen, V.: A trigram HMM-based POS tagger for Indian languages. In: Satapathy, S.C., Udgata, S.K., Biswal, B.N. (eds.) Proceedings of the International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA). AISC, vol. 199, pp. 205–212. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-35314-7_24
25. Smit, P., Virpioja, S., Grönroos, S.A., Kurimo, M.: Morfessor 2.0: toolkit for statistical morphological segmentation. In: Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics, pp. 21–24. Association for Computational Linguistics, Gothenburg (2014)
26. Srivastava, P., Chauhan, K., Aggarwal, D., Shukla, A., Dhar, J., Jain, V.P.: Deep learning based unsupervised POS tagging for Sanskrit. In: Proceedings of the 2018 International Conference on Algorithms, Computing and Artificial Intelligence, ACAI 2018. Association for Computing Machinery, New York (2018)
27. Sunitha, K.N., Kalyani, N.: A novel approach to improve rule based Telugu morphological analyzer. In: 2009 World Congress on Nature & Biologically Inspired Computing (NaBIC), pp. 1649–1652 (2009)
28. Trail, R.L.: The grammar of Lamani. SIL of the University of Oklahoma (1970)
29. Yu, X., Vu, N.T., Kuhn, J.: Ensemble self-training for low-resource languages: grapheme-to-phoneme conversion and morphological inflection. In: Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology, pp. 70–78 (2020)