Language in India www.languageinindia.com ISSN 1930-2940 Vol. 25:6 June 2025

Maithili Language Technology: A Survey

Shantanu Kumar, Ph.D.
Central Institute of Indian Languages
Mysore
shantanuk.ciil@gmail.com

Narayan Choudhary, Ph.D.
Central Institute of Indian Languages
Mysore
nchoudhary.ciil@gmail.com

Abstract
Maithili is an Indo-Aryan language primarily spoken in the Indian states of Bihar, Jharkhand, and the lower Terai regions of Nepal. While there has been some recent progress in Maithili language technology, the field is still nascent, with limited resources and research, and there is still a long way to go before we have robust and reliable language processing tools. More research and resources are needed to develop tools such as spell-checkers, text-to-speech systems, and machine translation models, which can help unlock the language's full potential in the digital age. This paper presents an effort made to survey the tools, systems, or research available in the field to facilitate language technology systems in Maithili. In this paper, the technological development of Maithili has been classified under certain heads, and the research works completed, along with the ones being undertaken, have been discussed. Different language technology systems, including the corpus creation and development presented for other languages, and their status for Maithili, have been investigated in this paper.

Keywords: Maithili, Language Technology, Corpus, ASR, TTS, MT, low-resource languages

Introduction

Maithili is an Indo-Aryan language primarily spoken in the Indian states of Bihar, Jharkhand, and parts of Nepal. While there has been some recent progress in Maithili language technology, the field is still nascent, with limited resources and research.

One of the significant challenges in developing Maithili language technology is the lack of a standardized orthography. The language has been written using a variety of scripts, including

Devanagari, *Kaithi*, and *Tirhuta*. This lack of standardization makes it difficult to develop tools such as spell-checkers and automated text-to-speech systems. Despite these challenges, some work has been done, discussed in the ensuing sections of the paper, in developing resources for Maithili language processing.

Although some progress has been made in Maithili language technology, there is still a long way to go before we have robust and reliable tools for processing the language. More research and resources are needed to develop tools such as spell-checkers, text-to-speech systems, and machine translation models, which can help to unlock the full potential of the language in the digital age.

While some commendable work has been done in Maithili language technology, there is still a lacuna in developing robust language processing tools for this language. The lack of standardized resources and corpora is one of the major challenges that must be addressed for further progress.

In this paper, the technological development of Maithili has been classified under certain heads, and the research works completed, along with the ones being undertaken, have been discussed below:

1. Corpus and Corpus Management

Corpus means a body of huge data incorporating various types of material, including text and audio-visual content in digital format. A corpus represents all the styles of a language, attempting to be as representative as possible for the goal it is designed. Corpus linguistics deals with the linguistic analysis of data available in a language. In the current scenario, when everything is supposed to be automatic, the need for text as well as speech corpora is indispensable. Any technology development process is based on the corpora available in that particular language. Therefore, it can be said that corpus and corpus management are the soul of language technology.

1.1 Tagged Corpus, Aligned Corpus, and Parallel Corpus

Corpora are classified into three types, namely Tagged Corpus, Parallel Corpus, and Aligned Corpus. A tagged corpus stands for a body of text or speech corpus tagged or annotated for a specific task, such as parts of speech, named entities, speech annotation, etc. Parallel corpora consist of corpora in two or more languages at the same level, that is, the engaging languages (minimum two) will have the same type of bilingual corpora available for both languages. Aligned corpora denote the property of corpora matched in both languages at every segment. It's a bilingual corpus aligned on a micro level in both languages.

1.2 Text Corpora

1.2.1 LDC-IL Maithili Text Corpora

LDC-IL Maithili Text Corpus (Choudhary, 2019) has been developed keeping in view the quality of the text, representativeness, retrievable format, size of corpus, authenticity, etc. For collecting a text corpus, LDC-IL adopted a standard category list of various domains and a prior set of criteria. The corpus of Maithili text can be broadly classified as literary and non-literary texts. Huge amounts of literary texts are available in Maithili, however, the number of scientific texts available is very less. Thus, the LDC-IL attempts to develop a balanced text corpus of Maithili. After the data has been collected from books, magazines, and newspapers, it is verified to be true to the original texts and then warehoused.

The Maithili Text Corpus is encoded in a machine-readable form and stored in a standard format. The major encoding format used for the purpose is Unicode, and the data is stored in XML format. The data is embedded with metadata information. The corpus has been created from the contemporary text in typed and crawled methods.

LDC-IL has a text corpus of 53,16,552 tokens extracted from 499 titles. This corpus contains five domains: Aesthetics, Commerce, Mass media, Science and technology, and social sciences (Ramamoorthy, 2019). Again, the consortium released a gold standard raw text corpus volume II, which includes 8,11,680 Words extracted from 54 Titles. The data covers three domains, which are further classified into 21 Sub-categories (Kumar, 2025). The corpus has been stored in XML format. Access to the LDC-IL Maithili Text Corpora is available through the LDC website. Interested researchers can apply for access by following the consortium's guidelines and procedures.

1.2.2 JNU Maithili Text Corpora

Under the Shallow Parser Tools Project in 2009 at JNU[1], a sizable corpus of Maithili was created. This corpus contains resources from various domains such as cuisine, astrology, history, language and literature, medical, politics, sports, tourism, and a small amount of information about the Mithila region regarding geography, history, and culture.

Though during the SwiftKey consultancy project, a huge amount of Maithili text corpora was created, which includes a frequency dictionary, a language model, and script input software, it is not an open-source dataset available for use.

1.2.3 Amity University Monolingual Corpora

A monolingual corpus of 25000 Maithili sentences (Nidhi, 2018) was created for the purpose of the development of Maithili-English statistical Machine translation systems. The monolingual Maithili data was collected for training the translation model in the target language.

1.2.4 Amity University Parallel Corpora

[1] http://sanskrit.jnu.ac.in/maithili/index.jsp

A Maithili-English divergence marked parallel corpora (Nidhi, 2018) was generated at the Amity University, Noida. They created a corpus of 45000 sentences and have given detailed documentation on the text corpora, alignment and parallelization strategies, training, testing, and a study of divergence between the language pair.

1.2.5 IIT-BHU Text Corpus

A team of researchers at the NLPR Lab, IIT-BHU, worked and released a text corpus. The data was created under Project Varanasi at the NLP Research Lab (NLPRL) at IIT (BHU), Varanasi, India, by a team of researchers and annotators. The Maithili Text Corpora[2] has Parts of Speech tagged 217887 Tokens, Chunked Corpus 1968 Sentences, Morphological Paradigm 2254 paradigms (Mundotiya, 2020).

1.2.6 Bhashini Parallel Corpora

Bhashini has released a corpus of 22535 Sentences aligning Maithili with English. The details of datasets such as parallel corpus, glossary corpus, or transliteration dataset, etc, can be found on the ULCA platform[3].

1.3 Speech Corpora

Speech Corpora refers to a specialized database that contains spoken audio files and their corresponding text, i.e., the written version of the spoken words or sentences. Recordings of people speaking. Speech corpora can be richly annotated with various linguistic information, such as phonetic transcription, prosodic annotation, speaker information, etc.

Speech corpora are crucial for various fields, particularly in:

- Speech Technology:
    - Automatic Speech Recognition (ASR): Training acoustic models that allow computers to understand spoken language.
    - Speaker Identification/Verification: Building systems to recognize or verify who is speaking.
    - Speech Synthesis: Creating artificial speech that sounds natural.
- Linguistics:
    - Phonetics and Phonology: Studying the sounds of language.
    - Conversation Analysis: Analyzing how people interact in spoken communication.
    - Dialectology: Researching regional variations in language.
    - Sociolinguistics: Studying the relationship between language and society.

[2] https://github.com/singhakr/Bhojpuri-Magahi-and-Maithili-Linguistic-Resources/blob/main/README.md
[3] https://bhashini.gov.in/ulca/dashboard

In this section, various types of speech corpora available for Maithili are discussed.

1.3.1 LDC-IL Maithili Raw Speech Corpora

The LDC-IL has created a speech corpus (Choudhary, 2019) for Maithili by covering the multiple geographic dialects of Sotipura, Bajjik, and Angika. The data has been collected from males and females of different age groups. This dataset consists of around 79 hours of audio with 45,198 audio segments. This dataset has a total of 306 speakers, with 150 males and 156 females. This dataset has been prepared to match the standard features for ASR datasets, such as 48 kHz sampling rate and 16 kbps bit rate (Ramamoorthy, 2019).

The speech corpora covers various domains such as contemporary text (news), creative text, sentences, date format, command and control words, person names, place names, most frequent words, phonetically balanced words, and form and function words.

The dataset released as volume II expands the raw speech corpus of Maithili. This dataset spans a duration of 109:09:50 (hh:mm:ss), consisting of read speech with continuous text, and spontaneous speech along with its transcription in Devnagari (Kumar, 2025). The data is derived from 49 female and 73 male native Maithili speakers, encompassing diverse age groups and regions (Rejitha, 2025). The data preparation follows the standard guidelines set by the consortium. Furthermore, the dataset includes the Angika variety to foster the corpus development for the language and its varieties. A comprehensive explanation of the dataset can be found in the Maithili Raw Speech Documentation at the data portal.

Researchers and language technology developers can use the corpus to study various aspects of Maithili language speech, such as its phonetics, prosody, and intonation. It can also be used to develop speech recognition and speech synthesis systems for Maithili.

1.3.2 LDC-IL Maithili Sentence Aligned Corpus

The LDC-IL has released a Sentence Aligned Speech Corpus suitable for ASR development in Maithili in 2023. The dataset comprises audio files in .wav format, accompanied by a corresponding textual layer containing phonetically normalized and orthographically normalized annotations in Devanagari script. This dataset spans a duration of 41:54:30 (hh:mm:ss), consisting of read speech with continuous text, representative sentences, and date formats (Kumar, 2023). The data consist of audio recordings from 147 female and 153 male native Maithili speakers, covering diverse age groups and regions. A comprehensive explanation of the dataset can be found in the Maithili Sentence Aligned Speech Documentation (Rejitha, 2023).

1.3.3 Dhwani speech corpora

Dhwani (Javed, 2022) is an unlabeled ASR corpus obtained from YouTube and News On AIR news bulletins. It has 37 hours of Maithili speech corpora. This corpus has been developed by the AI4Bharat team.

1.3.4 Vakyansh dataset

Vakyansh group, run by the EkStep foundation, has 56 hours of dataset (Chaddha, 2022), which has been collected from the AIR News bulletins. They collected the PDF files of the text and made a forced alignment for audio-to-text alignment. Thus, they have a wav2vec-based ASR model for Maithili. Due to automatic noise injection, there is a problem in their dataset that results in low-performing ASR systems. As a part of corpus development, the foundation has released the ASR dataset corpus of 140 hours (Labeled), 119 hours (Unlabeled), which is publicly available on GitHub[4].

1.3.5 IISc Speech Dataset

The Respin Project of IISc Bangalore released a dataset of 723:03:28 (hh:mm:ss) hours under the name of RESPIN Corpus(Saurabh 2025). The available corpus is part of a huge corpus of 10k hours of read-speech, which has 7.7M audio files and 197k sentences spoken by 18.8k speakers from 1500+ pincodes covering 9 Indian languages and 38 Dialects. It has 2 Domains: Agriculture & Finance. The details can be found on the Respin website[5].

1.4 Transliteration Dataset

1.4.1 AI4Bharat initiative

The AI4Bharat initiative, run by IITM, has a transliteration tool, Aksharantar (Madhani, 2022), which supports the Maithili language as well. For this tool, a dataset of Maithili has been prepared with the help of the StoryWeaver platform and involving the native Maithili speakers. The stories on the Storyweaver platform have been converted into tokens. This dataset consists of 21342 unique tokens and is available as an open-source dataset (Kumar, 2023).

1.4.2 Bhashini Transliteration Dataset

Bashini dataset, consisting of 221866 Sentences[6] (English-Maithili) general domain, is available on the ULCA platform. It includes some datasets submitted by the AI4Bharat team.

1.5 NER Dataset

[4] https://github.com/Open-Speech-EkStep/ULCA-asr-dataset-corpus/blob/main/README.md#maithili-labelled-external-total-duration-is-40-hours

[5] https://spiredatasets.ee.iisc.ac.in/respincorpus

[6] https://bhashini.gov.in/ulca/dashboard

IIT-BHU has created a dataset (Mundotiya, 2020) for Named Entity Recognition in Maithili, Magahi, and Bhojpuri. This dataset consists of 9815 Sentences, 157468 Tokens, 23338 Types, and 19809 Named Entities under 10,000 sentences. This dataset has been annotated with 22 entity labels using the IIIT-H NER guidelines. The NER model trained using this dataset has reported the lower baseline F1-scores from the NER tool obtained by using Conditional Random Fields models are 93.33 for Maithili. And the Deep Learning-based technique (LSTM-CNNs-CRF) achieved 93.33 for Maithili.

A similar work on the NER system has been reported by a group of researchers at the BIT Mesra, Ranchi. Though the researchers manually annotated 200k words, the corpus is not available in the public domain. After training the system, the final system achieved an f-measure of 91.6% with 94.9% precision and 88.53% recall(Priyadarshi, 2021).

A dataset for Maithili NER was developed as a part of a master's dissertation at Banaras Hindu University (Kumar, 2020). This dataset consists of 157470 tokens out of a total of 9814 sentences taken. Among these tokens, 21412 named entities were tagged, of which there were 10683 ENAMEX, 1837 TIMEX, 1335 NUMEX (IIIT NER tagset), and 1567 language-specific entities.

1.6 Text-to-Speech Corpus

A Text-to-Speech (TTS) corpus is a specially designed dataset used to develop and train Text-to-Speech systems, which convert written text into spoken audio.

1.6.1 LDC-IL TTS Corpus

The LDC-IL has released the first of its kind dataset for Maithili TTS, which enables the language to be included in the club of languages that have resources available for TTS technology. It is a valuable resource for developing speech technology for the Maithili language. The corpus spans a significant duration of 30 hours, 59 minutes, and 20 seconds. It has a total size of 19.56 GB. It contains 32,260 audio segments of read-speech from two native Maithili speakers (01 female and 01 male) recorded in a studio setup, ensuring high-quality audio (Kumar, 2025).

1.6.2 IndiaVoice-R TTS Corpus

The AI4Bharat team, through the IndiaVoice-R corpora, developed a TTS dataset for Maithili. IndicVoices-R (IV-R), the largest multilingual Indian TTS dataset, has been derived from an ASR dataset, with 1,704 hours of high-quality speech from 10,496 speakers across 22 Indian languages(Sankar, 2024). The database consists of 81.77 hours of multilingual speech from Maithili. It has further been divided into 6.18 hrs of read speech and 75.59 hrs of extempore speech. It has 32483 utterances spoken by 627 speakers.

2. Text Editors and Word Processors

There have been a lot of efforts made in terms of text editors and word processors in Maithili. A text editor refers to a tool that functions for text processing for a specific task. Similarly, a word processor refers to a tool that works on words for processing purposes, such as a morphological analyzer, spell checker, etc.

2.1 PoS Tagger

Parts of Speech (PoS) taggers are systems to annotate tokens with their corresponding part of speech tags automatically.

2.1.1 POS Tagset

A POS tagset was defined and manually annotated a Maithili corpus containing 52,190 words was manually annotated (Priyadarshi, 2020). This dataset was collected from a large corpus such as the Wikipedia dump and other Maithili web portals. With different model training and implementing various feature sets. They managed to get an accuracy of 85.88% for PoS tagging.

2.1.2 Maithili POS Tagger (MPOST)

The Maithili POS Tagger (MPOST) was developed by Dr Saroj Kumar Jha under the guidance of Prof Girish Nath Jha at the Centre for Sanskrit and Indic Studies, JNU, in collaboration with MGAHU, Wardha. The system is available on the centre's website[7].

2.2 Morph Analyser

A morph analyser for Maithili based on a finite state transducer has been developed at the Tezpur Central University (Rahi, 2020). This Maithili morphological analyser has been created on the LDC-IL corpus of around 855,430 words for performance evaluation and found no instance of failure on inflectional form as long as the root belongs to the lexicon file.

Also, there has been some work initiated on this system by JNU and the University of Hyderabad, but no results have been found yet in this regard.

2.3 Morph Generator

A Vowel Ending Approach model (Jha, 2018) on the Maithili Morph generator has been developed and reported by the research team of Mahatma Gandhi Antarrashtriya Hindi University, Wardha. They reported an accuracy of 71-74% with a 400k-word corpus.

2.4 Crawler

[7] http://sanskrit.jnu.ac.in/mpost/index.jsp

No crawler has been developed so far for Maithili by any research team or institution in India, or at least not anything could be found that is in the public domain.

2.5 Stemmer

The first ever stemmer for Maithili has been developed at BIT Mesra (Priyadarshi, 2019). This system has been developed using 110 sentences containing 2000 words, out of which 822 words were unique words' corpus developed in-house by the team. The baseline system achieved an accuracy of 70.9 % without any rule-based. When the POS dependent Rule-based module was implemented, the accuracy climbed to 84.6%.

3. Dictionary Tools

Though there have been a few efforts made by different institutions, including CIIL and JNU, for a Maithili online dictionary, there has not been any work reported yet for the online dictionary.

4. Spell Checkers/ Grammar Checkers / Style Checkers

In the word analyzer or spell checker front, there has not been any work done so far. This area needs to be explored in terms of developing different spelling tools such as dictionaries, spell checkers, grammar checkers, etc.

5. Parsing Systems

Maithili still stands on the chunking stage, and a parser has not been developed yet for the language. A research team at the IIT-BHU has developed a corpus of chunked data of 12,310 sentences (Mundotiya, 2021).

The TDIL has prepared WordNet[8] for 19 Indian languages, but Maithili has not been included yet among those languages.

6. Machine Translation and Translation Tools

1. A Java app has been developed at the Amity University, Noida (Nidhi, 2020). This app was developed during the development of English-Maithili translation systems. For this translation system, an application was developed to create a parallel corpus of English-Hindi and Maithili to save time. The Maithili language had very few resources, so Devanagari (used for Hindi) was an equivalent for the Maithili to train the model for the English-Maithili translation pair.
2. A team at the IIT-BHU has been working on developing resources and creating parallel corpora for three low-resource languages, Maithili, Magahi, and Bhojpuri (Mundotiya, 2020). A set of 10,000 sentences from Hindi, as a source language, has been translated into the three less-resourced languages. The present author was also a

[8] https://tdil-dc.in/index.php?option=com_vertical&parentid=90&lang=en

part of the corpora creation team and evaluated the Hindi-Maithili translated texts. For this research, they proposed a novel improvement in *Generative Adversarial Networks* (GAN)-NMT by incorporating deep reinforcement learning-based optimized attention in the generator and a convolutional neural network in the discriminator.

3. A translation tool, *Anuvadika*, has been developed by the researchers at CIIL[9]. This tool supports all 22 languages for English to Indic language translation. Maithili, supported by the Anuvadika tool, has shown a significant amount of accuracy with the translation.

7. Optical Character Recognition (Ocr)

Maithili is one of the languages that is not supported much by different tools in terms of optical character recognition. Not much work has been reported so far in the area of Maithili OCR.

Pramukh OCR[10] is a mobile application that supports twenty Indian languages, including Maithili. Though there are several tools and software that support Devanagari script for OCR but those are usually for Hindi; Maithili is reported as untouched in terms of a dedicated OCR tool.

8. Search Engines /Web Technologies

Shoonya[11] is an open-source platform to improve the efficacy of language work in Indian languages with AI tools and custom-built UI interfaces and features. This application has been developed by the AI4Bharat team at IIT-Madras. Shoonya supports various features for all 22 scheduled languages of India, and various tasks such as translation, text validation, speech transcription, optical character recognition, etc, can be done with this application.

9. Speech Technology

1. By the Vakyansh group, the Maithili ASR system has been developed along with around 40 other low-resource languages (Chaddha, 2022). They trained their model on 56 hours of the AIR dataset and reported a 12.8 % word error rate (WER) without a language model and 12.24% Word Error Rate (WER) with a language model.
2. The first ever work was reported on the Maithili Isolated Word Recognition system (Ranjan, 2016) at the JIIT Noida, New Delhi, using the HTK toolkit based on the Hidden Markov Model. He used 80 speech utterances and reported a 95% accuracy in recognition rate.
3. SpireLab[12] at the Indian Institute of Science, Bengaluru, is working on the collection of data for nine low-resource Indian languages, including Maithili. They are primarily

[9] https://anuvadika.ciil.org/
[10] https://www.pramukhocr.com/
[11] https://ai4bharat.iitm.ac.in/shoonya
[12] https://respin.iisc.ac.in/

working on the agriculture and finance domains. So far, they have not reported any work on ASR with their dataset.

4. A baseline Text to Speech system has been developed at the MGAHV, Wardha(Jha, 2019). For the system development, a total of 8 hours of audio data has been used, using both primary (collected from the field) and secondary data (collected from the LDC-IL speech corpus). This system reported an accuracy of 84 % when evaluated by the native speakers.

5. A significant research work has also been conducted at the CIIL for developing automatic speech recognition tools for Maithili with the dataset available with the LDC-IL. A tool named Maithili Anulekhika has been developed as a part of PhD work at the institute and is available for public use, which supports a live transcription facility of audio spoken in Maithili. Also, the transcription can be found by uploading any audio file on the web platform. This tool is the first of its kind that supports such a low-resource language, i.e., Maithili, and is freely available. The tool Maithili Anulekhika[13] can be accessed on Google engine (Kumar, 2024).

10. Text Input Technology

1. Predictive Keyboard[14] has been developed by the JNU research team. This keyboard is open source and can be downloaded on any mobile device.
2. Maithili is one of the languages available for wx-ITrans transliteration[15] at the JNU portal.
3. The Inscript Keyboard[16] Maithili (Mithilakshar) has been developed by the C-DAC team (2017). According to the Unicode standards, the Maithili script starts from 11480 and goes up to 114D9.

Apart from these researches in the area, there are a few more tools developed for Maithili text processing by the LDC-IL[17] team at CIIL, Mysore. A few of the tools have been mentioned below:

1. Frequency counter for consonants, vowels, the first letter in the word, and the sentence.
2. Frequency editor
3. KWIC and KWOC retriever
4. N-gram - Character level and Word level (unigram, bigram, and trigram)
5. Automatic Transliteration - Transliteration for Indian Languages to Roman and Devanagari to Mithilakshar.
6. Storage Interface for Speech dataset

[13] https://anulekhika.ldcil.org/maithili/
[14] http://sanskrit.jnu.ac.in/maithili/index.jsp
[15] http://sanskrit.jnu.ac.in/ile/index.jsp
[16] https://www.cdac.in/index.aspx?id=dl_mlingual_tools
[17] https://medha.ciil.org/

11. Standardization Issues

Despite being recognized as an official language of India, there are still standardization issues affecting its use and development. Mentioned below are some of the standardization issues in Maithili:

1. Lack of Standardization: Maithili does not have a standardized script, and there is no universally accepted standard for its orthography. This has led to spelling, pronunciation, and grammar inconsistencies across different regions where it is spoken.
2. Divergent dialects: Maithili has several dialects, each with distinct features, making it difficult to establish a unified standard. The dialects of Maithili spoken in different regions often differ significantly in terms of pronunciation, vocabulary, and grammar.
3. Historical challenges: Maithili has faced historical challenges, including the suppression of the language during the British colonial era, which has slowed its development and standardization.
4. Lack of Education: Education in Maithili is limited, and there are very few opportunities to study the language at the university level. This has led to a lack of trained professionals in the field of Maithili linguistics and a limited corpus of literature in the language.
5. Digitalization: With the advent of digital technologies, there is a need for standardization in Maithili typing and encoding to facilitate the development of digital resources in the language. However, there is currently no consensus on any one encoding standard. This hinders the creation of Maithili language resources.

Standardization issues in Maithili are a major challenge for the development and promotion of the language. However, efforts are being made to overcome these challenges and establish a standardized form of the language that can be used in education, literature, and digital communication.

Conclusion

As we have seen in the above discussion, Maithili is one of the languages that needs attention in most of the domains for language technology development. Though there are certain areas where researchers are working to develop language technology for the language, still, most of the areas in the field have been untouched, opening up a wide range for prospective researchers in the field. Fields such as Machine Translation, Parsers, Sentiment Analysis systems, NER systems, Anaphora resolution systems, WordNet system, ASR, TTS, OCR, spell checker, online dictionary, and many more are there, which opens up a diverse range to work on for the Maithili language.

References:

Choudhary, Narayan & L. Ramamoorthy. 2019. "LDC-IL Raw Text Corpora: An Overview" in *Linguistic Resources for AI/NLP in Indian Languages*, Central Institute of Indian Languages, Mysore. pp. 1-10.

Choudhary, Narayan, Rajesha N., Manasa G. & L. Ramamoorthy. 2019. "LDC-IL Raw Speech Corpora: An Overview" in *Linguistic Resources for AI/NLP in Indian Languages*. Central Institute of Indian Languages, Mysore. pp. 160-174.

IndicNLPSuite: Monolingual Corpora, Evaluation Benchmarks and Pre-trained Multilingual Language Models for Indian Languages. Kakwani, D., Kunchukuttan, A., Golla, S., Bhattacharyya, A., Khapra, M. M., & Kumar, P. (2020). Findings of EMNLP.

Javed, T., Doddapaneni, S., Raman, A., Bhogale, K. S., Ramesh, G., Kunchukuttan, A., ... & Khapra, M. M. (2022, June). Towards building ASR systems for the next billion users. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 36, No. 10, pp. 10813-10821).

Jha, A. K., Singh, P. P., & Dwivedi, P. (2019, July). The Maithili text-to-speech system. In 2019 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT) (pp. 1-6). IEEE.

Jha, S.K., Singh, P. P., & Kaul, V. K. (2018). VEA Model in Word Formation Process of Maithili MT.

Kumar, S.(2020). Named Entity Recognition in Maithili. Banaras Hindu University, Varanasi.

Kumar, A. Pratap and A. K. Singh, "Generative Adversarial Neural Machine Translation for Phonetic Languages via Reinforcement Learning," in *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 7, no. 1, pp. 190-199, Feb. 2023, doi: 10.1109/TETCI.2022.3209394.

Kumar, S., Mishra, D., Rajesha, N., Manasa, G., Srikanth, D., Fernandes, S., Nithin, S., Choudhary, N. K., & Mohan, S. (2023). *Maithili Sentence-Aligned Speech Corpus*. Central Institute of Indian Languages. ISBN 978-81-19411-96-2.

Kumar, S., Mishra, D., Varik, S., Fernandes, S., Nithin, S., Roopashri, M. R., Choudhary, N. K., & Mohan, S. (2025). The *Maithili Text-To-Speech Corpus*. Central Institute of Indian Languages. ISBN 978-93-48633-36-1.

Kumar, S., Tiwari, A., Rajesha, N., Manasa, G., Choudhary, N. K., & Mohan, S. (2025). *A Gold Standard Maithili Raw Text Corpus (Vol. II)*. Central Institute of Indian Languages. ISBN 978-93-48633-01-9.

Kumar, S., Tiwari, A., Rajesha, N., Manasa, G., Choudhary, N. K., & Mohan, S. (2025). *Maithili Raw Speech Corpus (Vol. II)*. Central Institute of Indian Languages. ISBN 978-93-48633-37-8.

Kumar, S. (2024). *Automatic speech recognition in Maithili: Issues & challenges* (Unpublished doctoral dissertation). Central Institute of Indian Languages.

Madhani, Y., Parthan, S., Bedekar, P., Khapra, R., Seshadri, V., Kunchukuttan, A., ... & Khapra, M. M. (2022). Aksharantar: Towards building open transliteration tools for the next billion users. *arXiv preprint arXiv:2205.03018*.

Mundotiya, R. K., Kumar, S., Chaudhary, U. C., Chauhan, S., Mishra, S., Gatla, P., & Singh, A. K. (2020). Development of a Dataset and a Deep Learning Baseline Named Entity Recognizer

for Three Low-Resource Languages: Bhojpuri, Maithili, and Magahi. *Transactions on Asian and Low-Resource Language Information Processing*, *22*(1), 1-20.

Mundotiya, R. K., Singh, M. K., Kapur, R., Mishra, S., & Singh, A. K. (2021). Linguistic resources for Bhojpuri, Magahi, and Maithili: Statistics about them, their similarity estimates, and baselines for three applications. *Transactions on Asian and Low-Resource Language Information Processing*, *20*(6), 1-37.

Mundotiya, R.K., Singh, M., Kapur, R., Mishra, S., & Singh, A.K. (2020). Linguistic Resources for Bhojpuri, Magahi, and Maithili: Statistics about Them, Their Similarity Estimates, and Baselines for Three Applications. *ACM Trans. Asian Low Resour. Lang. Inf. Process., 20*, 95:1-95:37.

Nidhi, R., & Singh, T. (2018). Resource Creation for English-Maithili Machine Translation (EMMT): A Divergence Perspective.

Nidhi, R., & Singh, T. (2018, August). English-Maithili machine translation and divergence. In 2018, 7th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO) (pp. 775-778). IEEE.

Nidhi, R., & Singh, T. (2020). Machine Translation and Divergence Study for English–Maithili. In *Information and Communication Technology for Sustainable Development: Proceedings of ICT4SD 2018* (pp. 383-390). Springer Singapore.

Nidhi, R., & Singh, T. (2020). SMT Algorithms for Indian Languages: Case Study of Moses and MT Hub for English-Maithili Language Pair. In *Proceedings of ICETIT 2019: Emerging Trends in Information Technology* (pp. 269-279). Springer International Publishing.

Priyadarshi, A., & Saha, S. K. (2019, December). A hybrid approach to developing the first stemmer in Maithili. In *Proceedings of the 11th Annual Meeting of the Forum for Information Retrieval Evaluation* (pp. 42-46).

Priyadarshi, A., & Saha, S. K. (2020). Towards the first Maithili part-of-speech tagger: Resource creation and system development. *Computer Speech & Language*, *62*, 101054.

Priyadarshi, A., & Saha, S. K. (2021). The first named entity recognizer in Maithili: Resource creation and system development. Journal of Intelligent & Fuzzy Systems, 41(1), 1083-1095.

Priyadarshi, A., & Saha, S. K. (2023). A study on the performance of Recurrent Neural Network-based models in Maithili Part of Speech Tagging. ACM Transactions on Asian and Low-Resource Language Information Processing, 22(2), 1-16.

Rahi, R., Pushp, S., Khan, A., & Sinha, S. K. (2020). A Finite State Transducer-Based Morphological Analyzer of Maithili Language. *arXiv preprint arXiv:2003.00234*.

Ramamoorthy, L., Narayan Choudhary, Arun Kumar Singh, Dinesh Mishra & Atuleshwar Jha. 2019. Maithili Raw Speech Corpus. Central Institute of Indian Languages, Mysore.

Ramamoorthy, L., Narayan Choudhary, Arun Kumar Singh & Dinesh Mishra. 2019. A Gold Standard Maithili Raw Text Corpus. Central Institute of Indian Languages, Mysore.

Ranjan, R., & Dubey, R. K. (2016, December). Isolated word recognition using HMM for the Maithili dialect. In *2016 International Conference on Signal Processing and Communication (ICSC)* (pp. 323-327). IEEE.

Rejitha, K. S., & Choudhary, N. K. (Eds.). (2023). *Compendium of LDC-IL Sentence-Aligned Speech Corpus*. Central Institute of Indian Languages. ISBN 978-81-19411-34-4.

Rejitha, K. S., & Choudhary, N. K. (Eds.). (2025). *LDC-IL Corpus Insights*. Central Institute of Indian Languages. ISBN 978-93-48633-33-0.

Saurabh et al.2025. 'RESPIN_S1.0 Corpus - A read speech corpus of 10000+ hours in dialects of nine Indian Languages'.

Sankar, A., Anand, S., Varadhan, P., Thomas, S., Singal, M., Kumar, S., ... & Khapra, M. (2024). Indicvoices-r: Unlocking a massive multilingual multi-speaker speech corpus for scaling indian tts. Advances in Neural Information Processing Systems, 37, 68161-68182.

Singh Chadha, H., Gupta, A., Shah, P., Chhimwal, N., Dhuriya, A., Gaur, R., & Raghavan, V. (2022). Vakyansh: ASR Toolkit for Low-Resource Indic Languages. *arXiv e-prints*, arXiv-2203.

Corresponding Author:

Shantanu Kumar, PhD

Central Institute of Indian Languages, Mysore

Karnataka-570006

Email: shantanuk.ciil@gmail.com

Contact: +91 9336054808