International Journal of Applied Mathematics
Volume 38 No. 10s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

A PARALLEL CORPUS FOR ADVANCING ENGLISH–SANTALI NEURAL MACHINE TRANSLATION

[*1]Sunil Kumar Sahoo, [2]Bhramara Bar Biswal, [3]Satya Ranjan Dash

[*1]Research Scholar, Department of Computer Science & Engineering, GIET University, Gunupur, Odisha, India. ORCID ID: 0009-0003-2805-9565
[2]Assistant Professor, Department of Computer Science & Engineering, GIET University, Gunupur, Odisha, India. ORCID ID: 0009-0007-2788-989X
[3]Associate Professor, KIIT University, Odisha, India. ORCID ID: 0000-0002-7902-1183

Email: sunil.sahoo@giet.edu[1], bhramarabarbiswal@giet.edu[2], sdashfca@kiit.ac.in[3]

Abstract

Machine Translation (MT) poses a significant challenge in developing language corpora for low-resource languages due to their minimal digital availability. Building such corpora is essential for preserving and promoting these languages. Santali, for instance, has very limited representation across online resources, and no proper translation tools including Google Translate exist for it. Developing a translation framework under such constraints is particularly difficult, as issues like low translation accuracy and heavy computational requirements arise. To overcome these limitations, the proposed MT system employs EnSanCorp, an English-Santali parallel corpus designed to facilitate Neural Machine Translation (NMT). EnSanCorp is created using multiple approaches, such as web-based parallel data extraction and optical character recognition (OCR) applied to scanned documents. The OCR-based method also demonstrates its usefulness for building corpora of other low-resource languages lacking online data. EnSanCorp currently contains 5,930 aligned sentences, 39,646 English tokens, and 39,936 Santali tokens, making it the most extensive English-Santali corpus available for research and non-commercial purposes. Evaluation results show that the Bilingual Evaluation Understudy (BLEU) scores for Statistical Machine Translation (SMT) and NMT vary across word and sentence levels: for word pairs, the scores are 0.04 (SMT) and 1.10 (NMT); for sentence pairs, 1.15 (SMT) and 7.20 (NMT). The overall BLEU scores achieved are 0.05 for SMT and 3.10 for NMT.

Keywords: Parallel Corpus, Neural Machine Translation, Statistical Machine Translation, Low-Resource Languages, Santali, EnSanCorp

1. Introduction

Machine Translation (MT) is an automatic method that changes one natural language into another. It is used in many fields such as business, education, technology, healthcare, and government services. For learners of Indian languages, MT has become very useful. India is a country with many languages. The Constitution of India recognizes 22 official languages. According to the 2011 Census, India has 122 major languages, 1,599 regional languages, 13 writing scripts, and about 720 dialects. Around 30 languages are spoken by more than one million people, while 122 languages are spoken by more than 10,000 people. About 20% of people in India understand English, but 80% do not. Languages differ from region to region, such as Gujarati in the west, Assamese in the east, Kashmiri in the north, and Tamil in the south. Other important Indian languages include Hindi, Bengali, Odia, Telugu, Kannada, and Urdu. In 2015, the Government of India started the "Digital India" program to provide better access to services through technology. However, many Indians still lack basic knowledge of Hindi and English. The Indian subcontinent has great diversity in people, culture, traditions, and languages. Some groups are tribal, while others are modern. Over time, languages, customs, and cultures have changed and mixed. The Santhal community is one such group that has spread across many parts of India and even to nearby countries like Bangladesh and Nepal.

Received: August 05, 2025

Volume 38 No. 10s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

The Santali language is very old and has a large vocabulary. It belongs to the Kherwali group of the Austric language family. Santali is spoken mostly in Jharkhand, Odisha, West Bengal, Assam, and Bihar. However, due to the language barrier, Santali-speaking people face difficulties in daily life, such as in markets, hospitals, and public places. There is no strong translation system available for the Santali language. Machine Translation for low-resource languages like Santali is difficult because there is not enough data available online. Unlike popular languages, Santali lacks large text datasets, writing tools, and proper dictionaries. This makes it hard to build accurate translation systems. Even when models are trained, the translation quality is low because of limited resources. To solve this, researchers use modern MT methods such as Statistical Machine Translation (SMT) and Neural Machine Translation (NMT). These methods need a parallel corpus (same text available in two languages). For Santali, a new corpus called EnSanCorp has been developed. Data for this corpus was collected from books, PDFs, websites, and scanned images using Optical Character Recognition (OCR). The data was cleaned, filtered, and grouped into words, phrases, and sentences. The EnSanCorp system combines SMT and NMT for English-to-Santali translation. It aims to improve communication in many areas such as schools, hospitals, and marketplaces, where Santali speakers face language problems.

The main contributions of this work are:
1. A new MT system for English-to-Santali translation is introduced for low-resource languages.
2. Data is collected from different sources such as books, websites, and scanned files using OCR.
3. The data is cleaned by removing duplicates and organized into useful forms.
4. EnSanCorp applies both SMT and NMT methods to provide better English-to-Santali translation.

The rest of the paper is arranged as follows:
- Section 2 gives an overview of MT systems, their performance, and limits.
- Section 3 explains the proposed MT system with figures.
- Section 4 presents the results with graphs and tables.
- Section 5 provides the main conclusion of the study.

2. Literature Survey

This part explains some important research work on Machine Translation (MT) in India, especially for low-resource languages.

Muskaan Singh et al. [27] built a system to translate Sanskrit into Hindi using verses from the Bhagavad Gita. They used a Deep Neural Network (DNN) with tokenization and CountVectorizer. The model gave good BLEU and WER scores, but it was not effective for short or simple sentences.

Sandeep Saini and Vineet Sahula [28] developed an NMT system with a shallow RNN and LSTM to translate six Indian languages: Hindi, Bangla, Tamil, Telugu, Urdu, and Malayalam. The model achieved a BLEU score of 18.215 for English–Hindi but faced problems with long sentences and small datasets.

Sahinur Rahman Laskar et al. [29] worked on Assamese–Bengali translation. They used a Seq2Seq RNN with attention and trained it on a parallel corpus. The system achieved BLEU scores of 7.21 (Assamese→Bengali) and 10.10 (Bengali→Assamese). However, the translations were not always accurate in both directions.

Saikiran Gogineni et al. [30] tested eight advanced models for six Indian languages: Hindi, Bengali, Gujarati, Malayalam, Tamil, and Telugu. Word embeddings and attention mechanisms were used. A four-layer Bi-LSTM achieved a BLEU score of 21.97. Short sentences were translated well, but longer ones were still a challenge.

Pushpalatha Kadavigere Nagaraj et al. [31] studied Kannada to English translation using an NMT system with LSTM in a Seq2Seq framework. The system performed well, reaching 86.32% accuracy and a good BLEU score, but the dataset used was small.

Received: August 05, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Kavit Gangar et al. [32] applied the Transformer model for Hindi–English translation. They used back-translation to expand the training data and Byte Pair Encoding (BPE) for tokenization. The best version of the model reached a BLEU score of 24.53. The drawback was that it required high computing power and more time.

Gaurav Tiwari et al. [33] compared two models, LSTM Seq2Seq and Convolutional Seq2Seq (ConvS2S), for English–Hindi translation. ConvS2S was faster and gave better BLEU scores. Using limited data, the model achieved a BLEU score of 16.28 with 8 layers.

Mani Bansal and D.K. Lobiyal [34] proposed a convolutional Seq2Seq model for English–Punjabi, Punjabi–English, Hindi–Punjabi, and Punjabi–Hindi. They used Gated Linear Units (GLU) and Multi-Hop Attention to handle context and dependencies. The method required high computational resources but gave low BLEU scores: 2 (English–Punjabi), 2.5 (Punjabi–English), 3.5 (Hindi–Punjabi), and 1.5 (Punjabi–Hindi).

From this survey, it is clear that MT systems face many problems such as low BLEU scores, heavy computation, and lack of large datasets. Some models work well for short sentences, while others struggle with long ones. Also, no proper method has been developed yet for English–Santali translation, which shows a big research gap.

Table 1: Survey of MT Techniques, Performance, and Limitations

Author & Reference	Technique Used	Limitations	Performance
Muskaan Singh et al. [27]	DNN-based corpus system	Weak for short sentences	39% WER
Sandeep Saini & Vineet Sahula [28]	Shallow RNN + LSTM	Problem with long sentences, small data	18.215 BLEU (English–Hindi)
Sahinur Rahman Laskar et al. [29]	Seq2Seq RNN with Attention	Not accurate in both directions	7.21 BLEU (Assamese→Bengali), 10.10 BLEU (Bengali→Assamese)
Saikiran Gogineni et al. [30]	Four-layer Bi-LSTM	Long sentences difficult	21.97 BLEU
Pushpalatha K. Nagaraj et al. [31]	Seq2Seq with LSTM	Limited data	86.32% Accuracy
Kavit Gangar et al. [32]	Transformer with BPE + Back-translation	High computation time	24.53 BLEU
Gaurav Tiwari et al. [33]	LSTM Seq2Seq vs. ConvS2S	Small dataset used	16.28 BLEU (8 layers)
Mani Bansal & D.K. Lobiyal [34]	ConvS2S with GLU + Multi-Hop Attention	High resources needed	BLEU: 2 (En–Pa), 2.5 (Pa–En), 3.5 (Hi–Pa), 1.5 (Pa–Hi)

3. Proposed methodology

The Machine Translation (MT) system has gained popularity for effectively translating low-resource languages. It helps people communicate easily in everyday situations, such as in markets, schools, and hospitals. Therefore, the proposed system focuses on developing an efficient English-to-Santali translation model. To build this system, data were collected from various sources, including online

International Journal of Applied Mathematics
Volume 38 No. 10s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

platforms, books, and PDFs. The text from PDFs was extracted using Optical Character Recognition (OCR) technology, and additional data were obtained from existing parallel corpora. The collected data were then organized into word pairs, sentence pairs, and word-sentence pairs. Next, the processed data were used in the EnSanCorp model, which combines both Statistical Machine Translation (SMT) and Neural Machine Translation (NMT) techniques to achieve accurate and fluent translations. The system's performance was evaluated using the BLEU score, considering both automatic and manual assessments of the English-to-Santali translation. Figure 1 illustrates the workflow of the proposed English-to-Santali translation model.

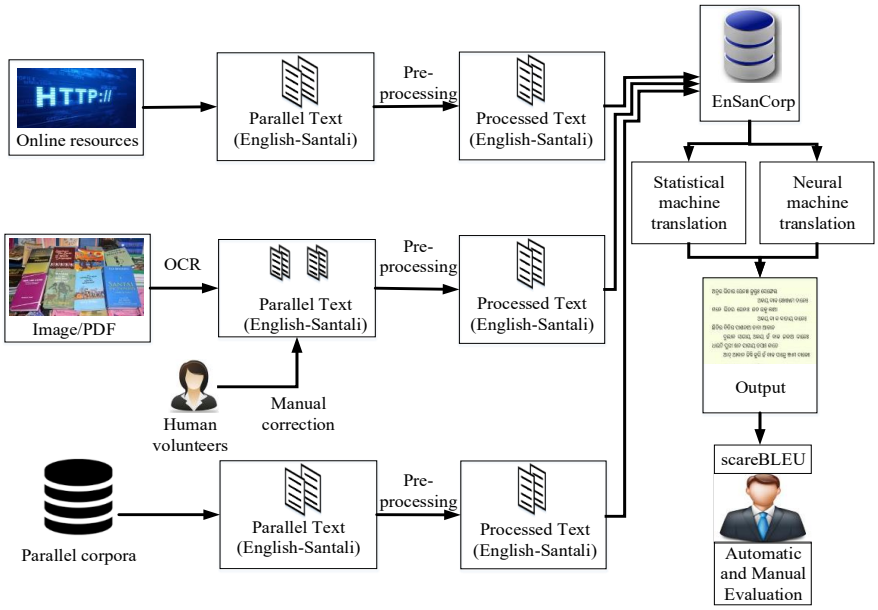


Figure 1: SMT and NMT Experimental Model

3.1 Data from Various Sources

Several online Santali resources are available, and different methods were explored to collect English–Santali parallel data for developing an efficient translation system. However, these approaches often require extensive manual processing to obtain a sufficiently large dataset.

The collected data were compiled by combining information from multiple sources, as described below:

- Text extracted using the **OCR technique**,
- Text gathered from **online resources**,
- Text **reused from existing corpora**.

Additionally, data were collected from several Santali language books, including *Fundamentals of Santali Words Book*, *Odisha Virtual Academy*, *A Concise English–Santali Book*, and *Sahaja Santali Sikshya Book*.

Assuming the dataset is composed of three main components, the total combined dataset can be expressed as the union of these three subsets. The mathematical representation of the total dataset is given as follows:

$$D_{total} = D_{OCR} \cup D_{Online} \cup D_{Corpus} \quad (1)$$

3.1.1 Text Extraction Done Based on Optical Character Recognition

OCR Code converts a two-dimensional image of text, which may contain machine-printed or hand-written text, into machine-readable text using the Python programming language [35]. The following is the OCR code.

Received: August 05, 2025

International Journal of Applied Mathematics
Volume 38 No. 10s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Proposed Algorithm
Step 1: Start of the program.
Step 2: Import required Python libraries:

- cv2 for image processing (OpenCV).
- numpy for numerical operations.
- pytesseract for OCR (Optical Character Recognition).
- PIL.Image for image handling.
- glob and os for file operations.

Step 3: Define a function get_string_Odia(img_path) that takes the image file path as input.
Step 4: Read the image using cv2.imread(img_path).
Step 5: Convert the image to grayscale using cv2.cvtColor(img, cv2.COLOR_BGR2GRAY) to simplify processing.
Step 6: Create a kernel matrix using np.ones((1,1), np.uint8) for morphological operations.
Step 7: Apply dilation and erosion using OpenCV:

- Dilation removes small black noises.
- Erosion helps remove white noise and improves character boundaries.

Step 8: Set the Tesseract configuration path using:
tessdata_dir_config = r'--tessdata-dir "C:\Users\Sunil\Desktop\OCR_Code\OCR_Code\sat"'

tep 9: Perform text extraction (OCR) from the processed image using:
pytesseract. image_to_string (img, lang="sat", config=tessdata_dir_config)

Step 10: Return the extracted text result.
Step 11: Call the function:
k = get_string_Odia("C:\Users\Sunil\Desktop\OCR_Code\OCR_Code\a.jpg") print(k)
Step 12: End of the program.

Extracting text using OCR has already been regarded as a main source of data collection in low-resource scenarios when web content is limited or non-existent. Extracting monolingual data using OCR helps improve MT performance for low-resource languages. OCR data extraction technology has improved substantially for large-scale textual resource digitalization. It is utilized to extract data from books, ancient hand-written documents, old newspapers, and so on. Sometimes, the OCR system makes mistakes in scanned texts as letters or falsely identifies text regions. It leads to OCR-Based linguistics and misspelling error in the output text.

3.1.2 Data gathered from several online resources

One of the challenging tasks is collecting the potential parallel texts from various websites and documents. For the websites related to Santali languages, there is very limited data presented, which are all explored to gather parallel data. Then, Crawl various websites related to Santali scripts and literature with a simple Python script. Web scraping using Python involves extracting data from websites using automated scripts. The following python code is the example of web scarp.
Step 1: Start
Step 2: Import Required Libraries

requests → for fetching web page content
BeautifulSoup → for parsing HTML

Received: August 05, 2025

International Journal of Applied Mathematics
Volume 38 No. 10s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

re → for cleaning text
os → for file handling

Step 3: Define the Target URL
url = https://example.com/santali-article
Step 4: Send HTTP Request
response = requests.get(url)
Step 5: Parse the HTML Content
soup = BeautifulSoup(response.text, "html.parser")
Step 6: Locate Santali Text Elements
santali_texts = soup.find_all('p')
santali_texts = soup.find_all('div', class_='santali-text')
Step 7: Extract and Clean Text
for t in santali_texts:
text = t.get_text()
Step 8: Store or Display Extracted Data
with open('santali_data.txt', 'a', encoding='utf-8') as f:
f.write(clean_text + "\n")
Step 9: Repeat for Multiple URLs (Optional)
Step 10: End

The majority of the data was gathered from Santali Wikipedia, which also covers information about the wide range of Santhal people and Santali culture. Some of the websites to collect the data as an online resource are given below.

- https://www.omniglot.com/writing/santali.html
- http://www.sdliit.ac.in/santali/santali_dictionary.pdf

These two websites, named SarjomBaha, include the Santali literature, news, and resources related to the Santali community. Then, Santali Online Dictionary which is an online dictionary for English to Santali transition. It also includes the definition of various words and phrases. The data was gathered from several online resources which involves searching and scraping the webpages. Let us assume the set of relevant web pages as P and the function of searching and extracting the text from these pages that is represented as S which is described as follows.

$$S : P \to D \quad (6)$$

Then, the extraction of a dataset from online resources is represented as,

$$T_{online} = \{S(p) \mid p \in P\} \quad (7)$$

3.1.3 Data extracted from existing Available Corpora

Some parallel data form various existing available corpora, such as the Santali language corpus, which is majorly utilized for linguistic research. The Santali language's word, token, and sentence counts are all included [36]. It also includes the lexicons, parts of speech tags, and lemmatized corpus based on nouns, prepositions, verbs, adverbs, adjectives, pronouns, conjunctions, and numerals. In this Santali corpus, nearly 590,314 tokens, 63,199 sentences, and 425238 words were included with some statistical information. The predetermined set of parallel or monolingual text data is extracted from existing corpora. Let us assume the set of available corpora as C which is described as follows.

International Journal of Applied Mathematics
Volume 38 No. 10s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

$$T_{corpora} = \bigcup_{c \in P} c \tag{8}$$

Here, each corpus is denoted as c which includes annotated linguistic data such as word counts,parts of speech, sentence counts, and so on.

3.1.4 Dataset composition

The data from these three sources are combined to ensure a comprehensive dataset which is described as,

$$T = T_{OCR} \cup T_{online} \cup T_{corpora} \tag{9}$$

Every element in T is a tuple (a,b), here the sentence or phrase in English is denoted as a and the corresponding translation in Santali language is denoted as t. The final dataset T process the cleaning and validation processs based on the possibility of errors.

$$T_{clean} = \{(a,b) \in T \mid validate(a,b)\} \tag{10}$$

Here, the function of validation is represented as $validate(a,b)$ for analyze the quality of data for linguistic consistency, noise removal, and correctness of translations. This process is performed to obtain a high-quality dataset.

3.2 Data processing

Aftercollectingtherawdata,asshowninTables2and3, in which the data from the corpus passes through various processes to provide the translated data. Then, it is categorized in to three groups based on the overall format which is described as follows.

- **Word pairs:** The word pairs are the samples that are primarily obtained from the word dictionary.
- **Sentence pairs:** The samples with a sequence in both the source and the target are called sentence pairs.
- **Word and sentence pairs:** This group is the concatenation of the previous two groups.

The earlier classification appears to indicate that a sample consists simply of word or sentence pairs. Sentences frequently appear within word pairs of data and vice versa. Next, the word and phrase pair data are split in the following ratio, 70:15:15, into train, development, and test sets.

Table 2: EnSanCorp word pairs along with their tokens

Source	**Tokens**	
	English	**Santali**
Fundamentals of Santali Word Book	17200	17360
Odisha Virtual Academy	1050	1060
A Concise English-Santali Book	4800	4900
Sahaja Santali Sikshya	6596	6616
Total	29646	29936

The word pairs of the data's token are described in Table 2 based on the resource of Santali language data collection. Each dataset collection was analyzed separately for both English and word pair data. It analyzed the data from fundamentals of the Santali language, Odisha Virtual Academy, Sahaja Santali Sikshya and the total source of the Santali language. EnSan Corp Sentence pair's parallel corpus, along with their token, is described in Table 3.

Received: August 05, 2025

International Journal of Applied Mathematics
Volume 38 No. 10s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Table 3: EnSanCorp Sentence pairs parallel corpus along with their tokens.

Source	Sentences (Parallel)	Tokens from English	Sentences (Santali)
Fundamentals of Santali Language	2299	9505	9597
Odisha Virtual Academy	270	2617	2637
Sahaja Santali Sikshya	40	425	430
Total	2609	12547	12664

The EnSanCorp Sentence pairs' parallel corpora are analyzed along with their English and Santali. It was performed for a parallel corpus, which included 270 sentences classified in the Odisha Virtual Academy and 2299 sentences in the core Santali language. Afterwards, a total of 40 sentences were gathered in Sahaja Santali Sikshya. Table 4 lists the total size for each dataset along with the sizes of the three stages, such as training, development, and test splits for different datasets.

Table 4: Sizes of training, development and test splits of various datasets with total size of each dataset.

Data	Train	Development	Test	Total size
Word pairs	20752	4447	4447	**29646**
Sentence pairs	2051	440	439	**2930**
Word and sentence pairs	22803	4887	4889	**32576**

Sizes of training, development, and test splits of various datasets with the total size of each dataset are described in Table 4 based on the Word pairs, Sentence pairs, and Word and Sentence pairs.

3.2.1Domain Coverage

The corpus that resulted in EnSan Corphas covered a wide variety of domains, especially when compared to corpora of a similar corpus. The corpus includes conversations in public places suchasmarkets, banks, tourism, and when rituals and cultural practices are observed.

3.3 Baseline of EnSanCorp

En San Corp MT system was developed to translate English and Santali languages. It is performed based on the combined SMT and NMT system.

3.3.1 SMT

The primary function of the SMT system is to mechanically map sentences from one human language (source) to another (target) [37]. It finds equation translation for the supplied language by utilizing a significant amount of bilingual data and probabilistic models to generate translation between languages. It is quite a successful architecture compared to other rule-based MT (RBMT). SMT also analyze the statistical and probabilistic way to analyze the information and conversion. For English–Santali language translation, the SMT model first employs the IBM model 2, which consists of an alignment model and a lexical translation model. It handles word reordering by introducing an alignment probability in addition to the word-level translation probability to simulate the absolute distortion in the word location between the source and target languages. First, the Moses tokenizer was used to tokenize the text in both the source and destination languages, with the language set to English. Next, nltk.translate.api's IBM translation model 2 should be trained. The Moses English de-tokenizer was used to perform the de-tokenization after acquiring the translation target tokens.

International Journal of Applied Mathematics
Volume 38 No. 10s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

3.3.2 NMT

The NMT system is a sophisticated translation method that predicts a word set by using neural network techniques [38]. It delivers accurate translation in a shorter amount of time, and it makes use of certain neural approaches to improve translation performance. The three most widely used neural network algorithms are Recurrent Neural Network (RNN), Long Short Term Memory (LSTM), and Gated Recurrent Units (GRUs).However, this technique has some issues, such as RNN having higher training time and low training speed. LSTM are computationally extensive, which reduces the training and evaluation speed. GRU are also computationally extensive, which makes it difficult to train large models, and it suffer from vanishing gradients. To address these concerns, the transformer is employed to train the dataset. NMT's translation model architecture is the transformer. The model was trained to translate English into Santali.

3.4 Proposed Methodology

Initially, sentence component BPE tokenization was taught by merging source and target language text. The maximum vocabulary size was set at 8k. The transformer model was built with PyTorch. The number of encoder and decoder levels was set at 3 with 8 heads in each layer. BPE is a subword tokenization method that enables the effective handling of morphological changes and phrases that are not part of the lexicon. The most frequent character combinations or character sequences in the training data are merged iteratively to develop the system.The self-attention layer's hidden dimension was set to 128, while the position-wise feed-forward layer's dimension was set to 512.The encoder and decoder both employed a dropout of 0.1. The encoder and decoder embedding layers were not interconnected. The model was trained with early halting and a patience of 5 epochs. Model training was halted if the validation loss did not improve for five consecutive epochs. Greedy decoding was utilized to create the translations during inference. Because the data was very tiny, the training and translation were carried out on the CPU.

3.4.1 Pre-processing

Initially, A joint vocabulary is generated by combining materials produced in the source and target languages. This ensures that there exist subwords in both languages, which is essential when translating between comparable or related languages.Each of the letters constitutes the fundamental vocabulary used by the BPE algorithm. Then, it combines the most common characters combined to create new tokens. Until a predetermined maximum vocabulary size is reached, this procedure remains progressing which is described as follows.

$$BPE(a) = \{a_1, a_2, ..., a_n\} \tag{11}$$

Here, the sentence is denoted as a and the BPE token is denoted as x_i. This BPE vocabulary is used to tokenize each sentence in the dataset, to obtain the sequence of subword units that are more manageable to train the neural network.

3.4.2 MTprocess

Transformers are specific types of neural network structures that function particularly well for language translation tasks since they depend on self-attention processes to handle long-range dependencies in sequences. There are three encoder and three decoder layers in the model. Every layer has a position-wise feed-forward network and a self-attention mechanism which are described as follows.

$$E_i = MHA(SA_i) + FF(SA_i)$$
$$D_i = MHA(SA_i) + MHA(E_i) \tag{12}$$

International Journal of Applied Mathematics
Volume 38 No. 10s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Here, MHA denoted Multihead attention which includeseight heads per layer which utilized to process the variousparts of the input sequence simultaneously. Then, SA_i denoted the self attention of i^{th} input sequence. The self-attention layers' hidden dimension is set to 128 while the position-wise feed-forward layer's dimension is set to 512. The size of the the model's internal representation and its ability to pick up intricate patterns are determined by these dimensions. Both the encoder and decoder layers use a dropout rate of 0.1 to randomly deactivate certain neurons during training in order to minimize overfitting. Since the encoder and decoder embedding layers are distinct, their weights are not shared. As a result, the model can acquire unique embeddings for both the target and source languages.Based on the input sequence, the model implies (translates) a sequence in the target language. Greedy decoding is a simple process that generates the final translation by selecting the most probable token at each stage.

4. Results and discussion

BLEU and chrF ratings were used to assess how well the translations in the test set performed. Both these scores were computed using the sacrebleu Python package. Sacre BLEU is a standardized implementation of the BLEU score, commonly used for evaluating the quality of machine translation models. For computing the BLEU score, set the to kenizertointlinsacrebleu to enableinternationaltokenizationbeforethecomputationoftheBLEUscore.The overall evaluation nresults of the proposed EnSan Corp technique are described briefly as follows.

4.1 Performance metric with its formulation

The overall performance of the proposed EnSan Corpwas analyzed using a particular formula to evaluate the efficiency. Some of the performance metrics utilized for the MT system, such as the BLEU score and ChrF score, are described as follows.

4.1.1 BLEU score

A performance statistic called the Bilingual Evaluation Understudy (BLEU) Score is used to compare a generated sentence to a reference sentence. It is computed as the N-gram that separates the MT system's output from the reference translation. Then, the precision value for n-grams of size 1 to 4 is computed and a brevity penalty for too short a translation. The mathematical expression for the BLEU score is described in the below equation.

$$BLEU = \min\left(1, \frac{OP\_len}{\mathrm{Re}f\_len}\right)\left(\prod_{i=1}^{4} pre_i\right)^{\frac{1}{4}} \tag{13}$$

Here, OP_len and $\mathrm{Re}f\_len$ are represented as the output length and reference length of the data. Here, pre denotes the precision value of n-gram.

4.1.2 ChrF score

Character n-gram F-Score ChrF is a performance metric utilized for automatic evaluation of MT output, which is described in the equation below.

$$ChrF\beta = \left(1+\beta^2\right)\frac{ChrP \cdot ChrR}{\beta^2 \cdot ChrP + ChrR} \tag{14}$$

Here, the character n-gram precision and recall are denoted as $ChrP$ and $ChrR$ which is averaged over all n-grams.

International Journal of Applied Mathematics
Volume 38 No. 10s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

4.1.3 ROUGE score

Recall-oriented understudy for Gisting evaluation (ROUGE) used to measure automatic summarization and quality of summary. The formula for ROUGE is described in the below equation.

$$ROUGE = \frac{no.of\ n\_grams\ found in \text{model and ref}}{no.of\ n\_grams\ in\ ref} \tag{15}$$

4.1.4 Word Error Rate (WER)

The WER is evlauted using Levenshtein distance which anlaysie the minimum count of word-level edits for translation which is described in below equation.

$$WER = \frac{A + C + F}{N} \tag{16}$$

Here, the number of incorrectly replaced words is denoted as A and the number of omitted words is denoted as C. The number of additionally added data is denoted as F and the total number of words are represented as N.

4.1.5 Translation Error Rate (TER)

The quality of MT is measured through TER metrics by evaluating number of modifications required for translation which is described in below equation.

$$TER = \frac{No.\ of\ \text{modification}}{total\ no.\ of\ words requied\ \ for translation} \tag{17}$$

4.2Manual evaluation of the proposed technique

In order to verify the automatic score, the proposed approach can access 300 randomly generated sentences from the SMT and NMT output. Figure 2 represents the loss curve for training and development while training the transformer model.

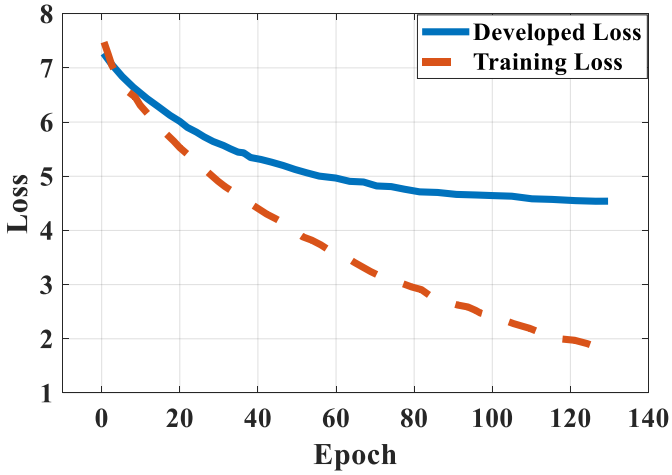


Figure 2: Training and developed loss curves while training the transformer model

The training and development loss were examined for the EnSanCorp technique for translating English to Santali while training the transformer model. The performance was evaluated using word pairs, sentence pairs, and word and sentence pairs.A machine learning model's performance is measured and evaluated using training and development loss, especially during the training stage. During the training phase, the model's performance on the training dataset is measured by the Training

International Journal of Applied Mathematics
Volume 38 No. 10s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Loss. The model's performance is measured using a different validation dataset that wasn't utilized during training, and this dataset is referred to as Development Loss or Validation Loss. The proposed technique achieves low training loss and limited developed loss. It describes the efficiency of the proposed EnSanCorp technique. Table 5 illustrates the manual evaluation samples and summary, respectively.

Table 5: Automatic evaluation results of MT

Data	**BLEU**		**CHRF**	
	Statistical MT	**Neural MT**	**Statistical MT**	**Neural MT**
Word pairs	0.04	1.10	0.10	0.10
Sentence pairs	1.15	7.20	0.21	0.30
Word and sentence pairs	0.05	3.10	0.11	0.20

The outcomes of the proposed EnSanCorp describe its efficiency for English to Santali language translation for SMT and NMT models. The BLEU and CHRF metric analyzed their performance for Word pairs, Sentence pairs, Word and Sentence pairs individually.It describes the efficiency of the proposed EnSanCorp technique for English to Santali corpus. Evaluation of BLEU score based on number of words and sample described in Table 6.

Table 6: BLEU score as a function of input sentence length for the test translation of sentence pairs data

Number of words	**Number of samples**	**BLEU**
1	3	3.25
2	37	3.19
3	124	6.27
4	113	7.89
5	50	5.13
6	43	4.01
7	21	6.60
8	14	4.30
9	13	11.31
10	7	3.61
11	6	3.92
12	4	14.85
13	1	7.67
14	1	1.48
17	1	0.90
19	1	0.00

The sample size and the number of English words translated were taken into account while calculating the BLEU score. These evaluations were utilized to measure the efficiency of English-to-Santali language translation. By comparing machine-generated translations with one or more reference translations, BLEU is used to assess the quality of such translations. A standardized version of BLEU named Sacre BLEU was developed to solve some of the problems and inconsistencies with the conventional BLEU evaluation process. To prepare for translations are shorter than the original translations, BLEU includes a brevity penalty and precision. The proposed method can only assess the BLUE score due to the small dataset. as it involves tokenization, n-gram selection, and brevity penalty, BLEU offers a significant degree of customisation as compared to the BLEU score.

4.3Analysis and discussion

The English-to-Santali statistical SMT and NMT systems have been tested many times with English-Santali parallel corpora. It has been observed that adding more parallel sentences to each domain parallel corpus can improve the translation quality. The length of the text given in Table 6 was used to calculate the BLEU score for the proposed EnSanCorp approach shown in Figure 3.

Figure 3: BLEU score modification based on the length of sentence

The evaluation of the BLEU score is determined by the length of each sentence separately. It was utilized to verify the efficiency of the proposed EnSanCorp technique for low-resource language translation (English Santali). The BLEU score is evaluated for varing number of length of sentence or number of words to efficient determine the efficient performance of proposed technique. The overall performance of the proposed EnSanCorp technique was analyzed based on several criteria named Excellent, Good, Partly correct, Ambiguity and incomplete, which is described in Table 7.

Table 7: Manual Evaluation Criteria

Excellent	No errors in the translation
Good	Translation are mostly accurate and complete but require a small correction.
Partly Correct	Part of the section is correct, but some words were mistranslated.
Ambiguity	Word's meaning was wrongly misunderstood by the MT system.
Incomplete	Parts are well-produced but conclude very early and missing some keywords.

The quality of the translation improves as the quantity of the proposed corpus increases.Then, the manual evaluation of the proposed EnSanCorp technique was analyzed for both English and Santali languages as input and output based on the BLEU score. Some of the Sample translations over the EnSanCorp technique are described in Table 8.

Table 8: Sample Manual evaluation for both English input and Santali outcomes

Manual Evaluation Result	English input	Santali output	Remarks
Excellent	They will go	[illegible]	Output conveying completer sentence
	I go	[illegible]	
	We are going	[illegible]	
Good	I have gone	[illegible]	Correct
	You have been eating	[illegible]	
Partially correct	We two will listen	[illegible]	Part of the sentence correct
	They have eaten	[illegible]	
Ambiguity	We are going	[illegible]	Misunderstood a word's meaning
	I will wait	[illegible]	
Incomplete	They will run	[illegible]	Missing some contents word
	Birds flew away	[illegible]	

International Journal of Applied Mathematics
Volume 38 No. 10s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

The Remarks are based on the manual evaluation of EnSanCorp for both English and Santali languages. They describe the efficiency level of EnSanCorp based on Excellent, Good, Partly Correct, Ambiguity, and Incomplete. Based on this concept, the manual evaluation of proposed techniques determines its efficient performance. The automatic and manual evaluation of EnSanCorp analyzes the quality of translation, represented in Figure 4.

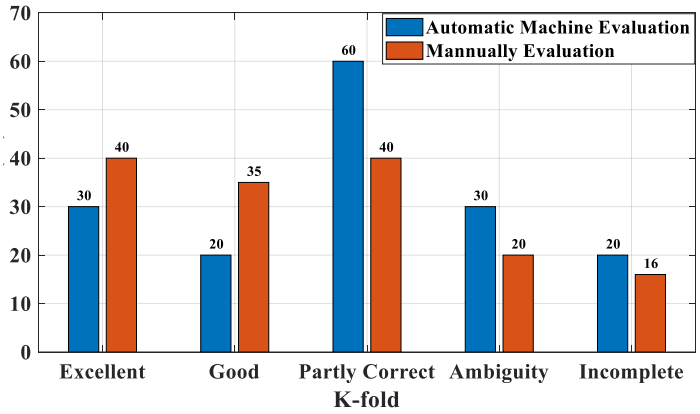


Figure4: Automatic and Manual evaluation summary of EnSanCorp

The overall evaluation summary for EnSan Corpis analyzed and represented in Figure 4 based on the manual evaluation. It describes the efficiency of EnSanCorp based on the combined SMT and NMT system.

4.4 Comparative analysis

The performance of the proposed MT system compared with other basis MT technique by anlaysis their WER, TER and ROUCH socore which is reprensted in the Figure 5.

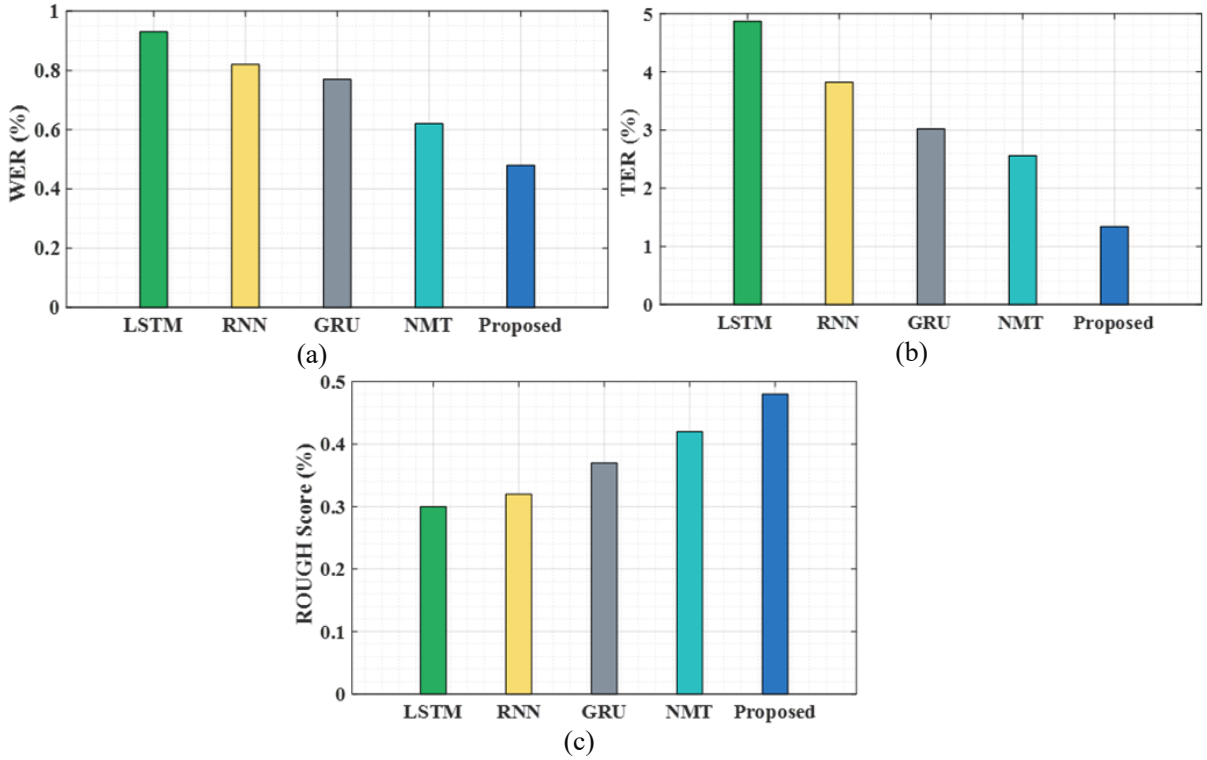


Figure 5: Performaneanlaysis of (a) WER, (b) TER and (c) ROUGH score

Received: August 05, 2025

International Journal of Applied Mathematics
Volume 38 No. 10s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

The performance of propsoed for WER, TER, and ROUGH scores are analyzed and compared with various related techniques such as LSTM, RNN, GRU, and NMT. The proposed technique achieves 0.48% which is higher than other existing low-resource MT system. Due to the lack of a Santali dataset, the performance of ROUGH is reduced. It will be improved in the future by constructing an efficient Santali dataset. Similarly, the error rate of the translation model is evaluated to determine the efficient performance of the proposed technique. The TER of the proposed technique is obtained to be 1.34 and the WER of this study is 0.4. This performance analysis represented the efficient translation of the proposed technique compared to other related techniques.

4.5 Case analysis and human evaluation

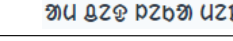
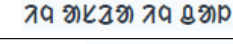
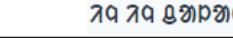
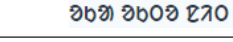
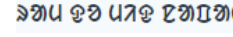
Evaluation of ROUGH, and TER score based on number of words and sample described in Table 9.

Table 9:ROUGH and TER score as a function of input sentence length for the test translation of sentence pairs data

Number of words	Number of samples	ROUGH	TER
1	3	0.27	2.27
2	37	0.21	4.34
3	124	0.27	5.27
4	113	0.89	7.89
5	50	0.13	10.13
6	43	0.01	15.01
7	21	0.68	10.68
8	14	0.36	9.36
9	13	0.98	9.98
10	7	0.56	12.34
11	6	0.92	15.53
12	4	0.65	12.45
13	1	0.44	13.34
14	1	0.56	20.43
17	1	0.32	17.34
19	1	0.32	18.00

The performance of ROUGH and ERT scores were analyzed in the proposed model with sample size and the number of English words translated was taken into account while calculating the BLEU score. These evaluations were utilized to measure the efficiency of English-to-Santali language translation. The human evaluation is conducted by translating the real-time data and their samples are described in Table 10.

Table 10: Samples of Human evaluation for English to Santali language translation

English input	Santali output	BLEU score
How are you?		7.56
I going to school		6.34
Shall I go		6.12
What is the time now?		5.56
Ram is good boy		6.98
Can you give me a pen		4.67

International Journal of Applied Mathematics
Volume 38 No. 10s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

The analysis of human evaluation describes the efficient performance of the proposed for translating English to the Santali language. It is determined by evaluating the BLEU score for each sentence of human-spoken data.

4.6 Discussion

The performance of EnSanCorpwas analyzed with several techniques based on the BLEU score. However, no related work has been done for the English-Santali parallel corpus. So, the comparison is conducted for EnSanCorp with other Indian low-resource language translation techniques based on SMT and NMT, which are described in Table 9.

Table 9: Comparison of some low-resource language translations with EnSanCorp

Author name and reference	**Technique used**	**Translation**	**BLEU Score**
SukantaSen et al. [39]	GRU and transformer network	Hindi-English and Hindi-Bengali	Approx. 1.38 – 15.36 points
HimanshuChoudhary et al. [40]	NMT using multihead self-attention with pre-trained BPE and MultiBPE embeddings.	English-Tamil and English-Malayalam	24.34 and 9.78
Aiusha V Hujon et al. [41]	NMT using transfer learning	English-Khasi language pair	39.63
Asha Hegde and HosahalliLaksmaiahShashirekha, [42]	KanSan Parallel Corpus (RNN, BiRNN and transformer-based NMT)	Kannada - Sanskrit and Sanskrit - Kannada	9.84 (Kannada – Sanskrit) and 12.63 (Sanskrit – Kannada)
Amit Kumar et al. [43]	Transfer learning-based Semi-supervised Pseudo-Corpus Generation approach	Bhojpuri-Hindi, Magahi-Hindi, Hindi-Bhojpuri and Hindi-Magahi	+15.56, +8.13, +3.98 and +2
Shweta Chauhan et al. [44]	Monolingual and Parallel corpora for Kangri low resource language	Kangri – Hindi translation	BLEU score (NMT – 3.25 and SMT – 4.98)
Proposed	**EnSanCorp based on SMT and NMT**	**English-Santali**	**3.19 obtained for 37 words**

SukantaSen et al. [31] created the Gated Recurrent Unit (GRU) and transformer network to improve the NMT in low-resource conditions without requiring any new data. To construct an efficient MT system, the NMT model employs multihead self-attention as well as pre-trained Byte-Pair-Encoded (BPE) and MultiBPE embeddings. It addresses the out-of-vocabulary (OOV) concerns for English-Tamil and English-Malayalam languages, as established by Himanshu Choudhary et al. [32].Then, MT system with transfer learning provides efficient English – Khasi language pairs developed by Aiusha V Hujon et al. [33], which provide a 39.63 BLEU score. Similarly, multiple techniques were analyzed using the BLEU score for automatic and manual English to Santali Corpus. Here, Table 9

Volume 38 No. 10s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

compares the BLEU scores of various related techniques over the Indian low-resource languages. It clearly describes that the proposed EnSanCorpachieves higher performance than other MT systems.

5. Conclusion

The English-Santali parallel corpus EnSanCorp, which is the framework for the proposed MT system, mainly focuses on NMT systems that could help translate English-Santali. EnSanCorp also uses optical character recognition (OCR) to extract parallel data from scanned images, as well as parallel data scraping from a variety of websites. For constructing a parallel corpus, the OCR-based data extraction method works well for especially low-resource languages with limited web content. A total of 29646 English words, 2930 phrases, and 29936 Santali tokens were in the EnSanCorp. The largest parallel English-Santali corpus covering a variety of subjects is called EnSanCorp, and it can be utilized free for academic and non-commercial uses. For four words and 113 samples, the BLEU score is 7.89; for nine words and thirteen samples, the BLEU score is 11.31. Similarly, the BLEU score was analyzed for the words (1, 2… 14, 17 and 19), which are analyzed with some amount of samples. EnSanCorp offers an effective translation solution for English to Santali language translation.Lack of Santail language dataset, where a limited amount to data was constructed which were gathered from various resource. In the future, advanced NMT approaches will be investigated to provide effective translation across the Santali language. More dataset resourcesrelated to the Santali language will be gathered to provide efficient Santali translation with high accuracy.

Compliance with Ethical Standards

Funding: No funding is provided for the preparation of manuscript.
Conflict of Interest: Authors declare that they have no conflict of interest.
Ethical Approval: This article does not contain any studies with human participants or animals performed by any of the authors.
Consent to participate: All the authors involved have agreed to participate in this submitted article.
Consent to Publish: All the authors involved in this manuscript give full consent for publication of this submitted article.
Authors Contributions: All authors have equal contributions in this work.
Data Availability Statement: Data sharing not applicable to this article.

References

[1] I. Rivera-Trigueros, Machine translation systems and quality assessment: a systematic review, Language Resources and Evaluation. 56(2) (2022) 593-619.
[2] N.A. Lone, K.J. Giri, and R. Bashir, Machine translation status of Indian scheduled languages: A survey. Multimedia Tools and Applications. (2023) 1–29.
[3] G. Huang, L. Liu, X. Wang, L. Wang, H. Li, Z. Tu, C. Huang, and S. Shi, Transmart: A practical interactive machine translation system, arXiv preprint arXiv:2105.13072. (2021).
[4] S.A.B. Andrabi, and A. Wahid, Machine translation system using deep learning for English to Urdu, Computational intelligence and neuroscience 2022 (2022).
[5] R. Vyas, K. Joshi, H. Sutar, and T.P. Nagarhalli, Real time machine translation system for english to indian language, In 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS). (2020) 838-842. IEEE.
[6] S. Parida, S. Panda, A. Dash, E. Villatoro-Tello, A.S. Doğruöz, R.M. Ortega-Mendoza, A. Hernández, Y. Sharma, and P. Motlicek, Open Machine Translation for Low Resource South American Languages (AmericasNLP 2021 Shared Task Contribution), In Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas. (2021)

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

218–223Association for Computational Linguistics, Online. https://doi.org/10.18653/v1/2021.americasnlp-1.24
[7] A. Santhanavijayan, D. Naresh Kumar, and G. Deepak, A novel hybridized strategy for machine translation of Indian languages, In Soft Computing and Signal Processing: Proceedings of 2nd ICSCSP 2019 2. (2020) 363-370.
[8] S. Dewangan, S. Alva, N. Joshi, and P. Bhattacharyya, Experience of neural machine translation between indian languages, Machine Translation. 35(1) (2021) 71-99.
[9] A. Kunchukuttan, and P. Bhattacharyya, Utilizing language relatedness to improve machine translation: A case study on languages of the indian subcontinent, arXiv preprint arXiv:2003.08925. (2020).
[10] S.K. Sahoo, B.K. Mishra, S. Parida, S.R. Dash, J.N. Besra, and E.V. Tello, Automatic Dialect Detection for Low Resource Santali Language. In 2021 19th OITS International Conference on Information Technology (OCIT). (2021) 234–238.
[11] S. Mahapatra, and I. Sarangi, Santhali Language in the Digital Media Space.
[12] T. Manjula, and T. Sudha, Recent development in speech recognition systems for Indian regional languages: A review.
[13] J. Nair, R. Ahammed, and A. Shaji, A Study on Transliteration Techniques and Conventional Transliteration Schemes for Indian Languages, In Sustainable Communication Networks and Application: Proceedings of ICSCN 2021. (2022) 103-117. Singapore: Springer Nature Singapore, 2022.
[14] J. Gala, P.A. Chitale, A.K. Raghavan, S. Doddapaneni, V. Gumma, A. Kumar, J. Nawale, A. Kunchukuttan, IndicTrans2: Towards High-Quality and Accessible Machine Translation Models for all 22 Scheduled Indian Languages, arXiv preprint arXiv:2305.16307. (2023).
[15] J. Basu, and S. Majumder, Performance evaluation of language identification on emotional speech corpus of three Indian languages, In Intelligence Enabled Research: DoSIER 2020. (2020) 55-63. Singapore: Springer Singapore.
[16] A. Gutkin, C. Johny, R. Doctor, L. Wolf-Sonkin, and B. Roark, Extensions to Brahmic script processing within the Nisaba library: new scripts, languages and utilities." In Proceedings of the Thirteenth Language Resources and Evaluation Conference. (2022) 6450-6460.
[17] B. AnsumaliMukhopadhyay, Ancestral Dravidian languages in Indus Civilization: ultraconserved Dravidian tooth-word reveals deep linguistic ancestry and supports genetics, Humanities and Social Sciences Communications. 8(1) (2021) 1-14.
[18] M. Aklin, B. Blankenship, V. Nandan, and J. Urpelainen, The great equalizer: Inequality in tribal energy access and policies to address it, Energy Research & Social Science. 79 (2021) 102132.
[19]Tonja, AtnafuLambebo, Olga Kolesnikova, Alexander Gelbukh, and GrigoriSidorov. "Low-resource neural machine translation improvement using source-side monolingual data." Applied Sciences 13, no. 2 (2023): 1201.
[20]Tonja, AtnafuLambebo, Olga Kolesnikova, Alexander Gelbukh, and GrigoriSidorov. "Low-resource neural machine translation improvement using source-side monolingual data." Applied Sciences 13, no. 2 (2023): 1201.
[21] Hendy, Amr, Mohamed Abdelrehim, AmrSharaf, VikasRaunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. "How good are gpt models at machine translation? a comprehensive evaluation." arXiv preprint arXiv:2302.09210 (2023).
[22] Ma, Shuming, Jian Yang, Haoyang Huang, Zewen Chi, Li Dong, Dongdong Zhang, Hany Hassan Awadalla et al. "Xlm-t: Scaling up multilingual machine translation with pretrained cross-lingual transformer encoders." arXiv preprint arXiv:2012.15547 (2020).
[23] Baziotis, Christos, Barry Haddow, and Alexandra Birch. "Language model prior for low-resource neural machine translation." arXiv preprint arXiv:2004.14928 (2020).

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

[24] Haque, Rejwanul, Chao-Hong Liu, and Andy Way. "Recent advances of low-resource neural machine translation." Machine Translation 35, no. 4 (2021): 451-474.
[25] Sen, Sukanta, Mohammed Hasanuzzaman, Asif Ekbal, Pushpak Bhattacharyya, and Andy Way. "Neural machine translation of low-resource languages using SMT phrase pair injection." Natural Language Engineering 27, no. 3 (2021): 271-292.
[26] Y. Bablani, S. Uqaili, S. Narejo, and H. Zahra, Survey on Text to Text Machine Translation, In 2nd International Conference on Computational Sciences and Technologies. (2020) 17-19.
[27] M. Singh, R. Kumar, and I. Chana, Corpus based machine translation system with deep neural network for Sanskrit to Hindi translation, Procedia Computer Science. 167 (2020) 2534-2544.
[28] S. Saini and V. Sahula, Setting up a neural machine translation system for English to Indian languages, In Cognitive Informatics, Computer Modelling, and Cognitive Science. (2020) 195-212. Academic Press.
[29] S.R. Laskar, P. Pakray, and S. Bandyopadhyay, Neural machine translation: Assamese–Bengali, In Modeling, Simulation and Optimization: Proceedings of CoMSO 2020. (2021) 571-579. Springer Singapore.
[30] S. Gogineni, G. Suryanarayana, and S.K. Surendran, An effective neural machine translation for english to hindi language, In 2020 International Conference on Smart Electronics and Communication (ICOSEC). (2020) 209-214. IEEE.
[31] P.K. Nagaraj, K.S. Ravikumar, M.S. Kasyap, M.H.S. Murthy, and J. Paul, Kannada to English Machine Translation Using Deep Neural Network, Ingénierie des Systèmes d Inf. 26(1) (2021) 123-127.
[32] K. Gangar, H. Ruparel, and S. Lele, Hindi to english: Transformer-based neural machine translation, In International Conference on Communication, Computing and Electronics Systems: Proceedings of ICCCES 2020. (2021) 337-347. Springer Singapore.
[33] G. Tiwari, A. Sharma, A. Sahotra, and R. Kapoor, English-Hindi neural machine translation-LSTM seq2seq and ConvS2S, In 2020 International Conference on Communication and Signal Processing (ICCSP). (2020) 871-875. IEEE.
[34] M. Bansal, and D. K. Lobiyal, Multi-lingual sequence to sequence convolutional machine translation, Multimedia Tools and Applications. 80(25) (2021) 33701-33726.
[35] K.C. Shekar, M.A. Cross, and V. Vasudevan, Optical character recognition and neural machine translation using deep learning techniques, In Innovations in Computer Science and Engineering: Proceedings of 8th ICICSE. (2021) 277-283. Springer Singapore.
[36] A.K. Akhtar, G. Sahoo, and M. Kumar, Digital corpus of Santali language, In 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI). (2017) 934-938. IEEE.
[37] H. Phan, and A. Jannesari, Statistical machine translation outperforms neural machine translation in software engineering: why and how, In Proceedings of the 1st ACM SIGSOFT International Workshop on Representation Learning for Software Engineering and Program Languages. (2020) 3-12.
[38] S. Sen, M. Hasanuzzaman, A. Ekbal, P. Bhattacharyya, and A. Way, Neural machine translation of low-resource languages using SMT phrase pair injection, Natural Language Engineering. 27(3) (2021) 271-292.
[40] H. Choudhary, S. Rao, and R. Rohilla, Neural machine translation for low-resourced Indian languages, arXiv preprint arXiv:2004.13819. (2020).
[41] A.V. Hujon, T.D. Singh, and K. Amitab,Transfer learning based neural machine translation of english-khasi on low-resource settings, Procedia Computer Science. 218 (2023) 1-8.
[42] A. Hegde, and H.L. Shashirekha, KanSan: Kannada-Sanskrit Parallel Corpus Construction for Machine Translation, In International Conference on Speech and Language Technologies for Low-resource Languages. (2022) 3-18.

International Journal of Applied Mathematics
Volume 38 No. 10s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

[43] A. Kumar, R.K. Mundotiya, A. Pratap, and A.K. Singh, TLSPG: Transfer learning-based semi-supervised pseudo-corpus generation approach for zero-shot translation, Journal of King Saud University-Computer and Information Sciences. 34(9) (2022) 6552-6563.
[44] S. Chauhan, S. Saxena, and P. Daniel, Monolingual and parallel corpora for Kangri low resource language, arXiv preprint arXiv:2103.11596 (2021).

Received: August 05, 2025