Evaluating IndicTrans2 and ByT5 for English–Santali Machine Translation Using the Ol Chiki Script

Kshetrimayum Boynao Singh[1,2], **Asif Ekbal**[1], **Partha Pakray**[2]
[1]Indian Institute of Technology Patna, India
[2]National Institute of Technology Silchar, India
boynfrancis@gmail.com, asif@iitp.ac.in, partha@cse.nits.ac.in

Abstract

In this study, we examine and evaluate two multilingual NMT models, IndicTrans2 and ByT5, for English-Santali bidirectional translation using the Ol Chiki script. The models are trained on the MMLoSo Shared Task dataset, supplemented with public English-Santali resources, and evaluated on the AI4Bharat IN22 and Flores test sets, specifically IN22-Gen and Flores200-dev. IndicTrans2 finetune strongly outperforms ByT5 across both directions. On IN22-Gen, it achieves 26.8 BLEU and 53.9 chrF++ for Santali→English and 7.3 BLEU and 40.3 chrF++ for English→Santali, compared to ByT5's 5.6 BLEU and 30.2 chrF++ for Santali→English and 2.9 BLEU and 32.6 chrF++ for English→Santali. On the Flores test set, IndicTrans2 finetune achieves 22 BLEU, 49.2 chrF++, and 4.7 BLEU, 32.7 chrF++. Again, it surpasses ByT5. While ByT5's byte-level modelling is script-agnostic, it struggles with Santali morphology. IndicTrans2 benefits from multilingual pre-training and script unification.

1 Introduction

Natural Language Processing (NLP) has made significant progress in a short amount of time, resulting in substantial improvements in machine translation (MT), particularly for languages with extensive resources. However, many low-resource languages (LRLs), particularly those spoken by indigenous and tribal groups, are underrepresented in digital spaces. Santali, a primary tribal language spoken by millions of people in India, Bangladesh, and Nepal, is a good example of this difference. Santali is widely spoken and essential to the culture, but it is still not well represented in digital form, with few language resources available for computational uses, such as machine translation.

A major problem with making strong machine translation systems for Santali is that there aren't many large, high-quality parallel corpora available (lrl, 2025) Conventional machine translation systems, including neural architectures, necessitate extensive bilingual data to proficiently acquire the mapping between source and target languages. The absence of annotated resources significantly hinders the translation quality for Santali, which employs the Ol Chiki script, a writing system that is not extensively supported in conventional NLP tools and tokenisers.

This research addresses these issues by evaluating two sophisticated multilingual translation models, IndicTrans2 (Gala et al., 2023) and ByT5 (Xue et al., 2022), for the bidirectional translation of English–Santali . IndicTrans2 is a transformer-based model that works best with Indian low-resource fo Indic languages (Pakray et al., 2024). It is well known for its ability to transfer information between languages in environments with limited resources. ByT5 works at the byte level, which means it can handle scripts and characters that aren't visible. This suggests that it could enhance language inclusivity and reach a diverse range of linguistic communities.

We use quantitative metrics, such as BLEU and chrF++, to evaluate the quality of translations into the Ol Chiki script. Our research aims to demonstrate the versatility of models for underrepresented languages and to promote the development of more inclusive and accessible machine translation systems for low-resource linguistic communities.

2 Related Works

Research in machine translation (MT) for low-resource (Singh et al., 2023b) languages (LRLs) has gained significant momentum in recent years, as the NLP community increasingly focuses on linguistic inclusivity and equitable access to technology. Early approaches to MT primarily relied on rule-based and statistical methods, which required extensive linguistic expertise and manually crafted translation rules. Although these systems were innovative for their time, they often suffered from scalability issues and produced suboptimal results for languages with limited parallel corpora (Singh et al., 2024b) and sparse digital resources.

The emergence of neural machine translation (NMT) (Appicharla et al., 2024) has revolutionised

Proceedings of MMLoSo 2025: First Workshop on Multimodal Models for Low-Resource Contexts and Social Impact, pages 95–100
December 23, 2025 ©2025 Association for Computational Linguistics

the field by introducing deep learning architectures capable of modelling complex language patterns. The Transformer model and its subsequent variants demonstrated remarkable improvements in translation fluency and accuracy, particularly for high-resource languages. However, NMT models remained heavily data-dependent, and their effectiveness diminished substantially for low-resource languages that lacked sufficient training data.

To address this limitation, researchers have begun developing multilingual NMT models (Singh et al., 2024a) capable of learning shared representations across multiple languages. Such models leverage (Singh et al., 2023a) cross-lingual transfer (Wei et al., 2024) learning, enabling knowledge gained from high-resource languages to improve translation quality for related low-resource ones. IndicTrans2 represents one such advancement tailored explicitly for Indian languages (Dabre and Kunchukuttan, 2024). It employs a shared multilingual (Limisiewicz et al., 2024) encoder–decoder architecture, effectively utilising linguistic similarities among Indo-Aryan and Dravidian languages to enhance performance for low-resource pairs.

In contrast, ByT5 belongs to a newer generation of models that operate at the byte level, eliminating the need for language-specific tokenisation. By processing text as raw bytes, ByT5 can seamlessly handle diverse writing systems and scripts, including those that are poorly represented in mainstream tokenisers. This makes it particularly advantageous for languages like Santali, which is written in the Ol Chiki script, a script not widely supported in traditional NLP pipelines.

Our work builds upon these foundational advances by applying and comparing IndicTrans2 and ByT5 to the English–Santali translation task. Through this comparative analysis, we aim to evaluate how multilingual and byte-level (Nehrdich et al., 2024) modelling approaches perform on a genuinely low-resource (Bhaskar and Krishnamurthy, 2024) language with unique orthographic and linguistic characteristics.

3 Linguistics of Santali

Santali[1] is one of the Munda languages in the Austroasiatic language family. Other Munda languages include Mundari, Ho, and Korku. The Santal community, which is one of the largest indigenous tribal groups in India, speaks it most of the time. The Santali-speaking population is estimated to be around 7 million, primarily residing in the Indian states of Jharkhand, West Bengal, Odisha, and Bihar. There are also smaller groups of speakers in Assam and neighbouring countries, such as Bangladesh and Nepal.

In the past, Santali was written in several scripts, including Devanagari, Bengali, Odia, and Latin-based orthographies. This was because of regional language influences and colonial legacies. In 1980, however, the Ol Chiki script, which Pandit Raghunath Murmu developed in the mid-20th century, was officially recognised as the standard writing system for Santali. Ol Chiki was made to better show the phonological structure of the language than borrowed scripts, which had problems with sound representation and spelling consistency.

The use of Ol Chiki has been very helpful in reviving Santali, maintaining its vitality, and making it more consistent. It has helped the language get more attention in formal education, literature, and online spaces. Alternative scripts are still used informally, particularly in regions with multiple languages. However, the Government of India now officially recognises Ol Chiki and supports it in the Unicode Standard, ensuring it works with modern computer systems.

To sum up, Santali is a vibrant language with numerous forms and rich cultural significance for the Santal people. The institutionalisation of Ol Chiki is a crucial step toward preserving its linguistic identity and ensuring it is represented in the digital and technological age.

4 Methodology

In this study, we delineate the methodology in two principal phases: dataset preprocessing and model architecture. First, we explain how we made and organized the English-Santali parallel dataset for training. Second, we examine the architecture of the two models used, IndicTrans2 and ByT5, and explain how each is utilised to translate from English–Santali.

4.1 Datasets

The dataset utilised in this study significantly expands upon the original 20,000 English–Santali parallel sentences provided by the MMLoSo 2025

[1] https://en.wikipedia.org/wiki/Santali_Wikipedia

Split	Sentence	English	Santali
Train Set	104,451	1,082,726	1,036,528
Valid	1503	14,850	16,001
Test IN22	1024	25,348	26,676
Test Flores	997	20,955	22,912
Total	**107,975**	**1,143,879**	**1,102,117**

Table 1: Dataset statistics for the English–Santali corpus. The table reports the number of sentences and total tokens in English and Santali across the training set, IN22-Conv validation set, IN22-Gen test set, and Flores200-dev test set.

Category	Details for ByT5 Model finetune
ByT5 Training Configuration	
Model	`google/byt5-small` (byte-level)
Batch Size	64
Learning Rate	3×10^{-4}
Epochs	5
Max Length	256 characters (source & target)
Evaluation Metrics	
Metrics	BLEU, chrF++, CER (Character Error Rate)
Important Flags	
`data_dir`	Path to dataset (required)
`output_dir`	Directory to save model (required)
`fp16`	Enable mixed precision training
`lowercase`	Normalize all text to lowercase

Table 2: Training configuration for ByT5 finetuning.

Category	Details for IndicTrans2 finetune
Optimization	
Optimizer	Adam ($\beta_1 = 0.9$, $\beta_2 = 0.98$)
Learning Rate	3×10^{-5} with inverse square root decay
Warmup	2,000 steps (from 1×10^{-7})
Gradient Clipping	1.0
Training Configuration	
Batch Size	32,768 tokens (effective across 8 GPUs)
Max Updates	100,000
Mixed Precision	FP16
Regularization	
Dropout	0.2
Label Smoothing	0.1
Early Stopping	Patience of 10 checkpoints (BLEU-based)

Table 3: Training configuration for IndicTrans2 finetuning.

organisers [2]. To enhance model robustness and improve translation quality, we incorporate publicly available English–Santali resources[3], followed by data augmentation techniques to increase corpus diversity further. After integration and preprocessing, the final dataset, as presented in Table 1, comprises 107,975 parallel sentences, covering a broad spectrum of domains, including culture, daily communication, news-style texts, and general knowledge.

All sentence pairs were manually inspected for alignment quality. Comprehensive preprocessing was applied, including text cleaning, Unicode normalisation for the Ol Chiki script, tokenisation, and consistency checks to ensure high-quality parallelism. The expanded corpus reflects Santali's rich morphology, with the Santali side containing 1,102,117 tokens, compared to 1,143,879 tokens on the English side.

The dataset is divided into a training set, IN22-Conv validation set, and two test sets (IN22-Gen and Flores200-dev). Table 1 provides the complete dataset breakdown. On average, English sentences contain 14–16 tokens, whereas Santali sentences contain 17–19 tokens, indicating Santali's morphologically richer structure. The dataset and related resources are available.[4]

4.2 Model Architecture

ByT5 Architecture: The proposed English-Santhali translation system is based on the ByT5 architecture Table 2, a token-free version of the T5 and mT5 (Xue et al., 2021) transformer models. It was designed to work directly on raw UTF-8 bytes, rather than using subword tokenisation. This design doesn't use word tokens, allowing the model to handle any language script, including low-resource and morphologically complex languages such as Santhali, which can be written in multiple scripts, like Ol Chiki and Devanagari. ByT5 has a small vocabulary of only 256 bytes. It uses a sequence of byte values to represent each character, which solves problems with words that aren't in the language, spelling mistakes, and noisy input. The model is based on an encoder-decoder framework, but it uses an unbalanced "heavy encoder" architecture, where the encoder is three times deeper than the decoder. The model can learn a soft lexicon by understanding word structure, morphology, and syntactic patterns directly from byte sequences thanks to this deeper encoder. The decoder, on the other hand, focuses on making coherent target text. Additionally, ByT5 relocates the parameters that traditional models utilise for large token embed-

[2] https://www.kaggle.com/competitions/mmloso2025/data?select=santali-train.csv

[3] https://opus.nlpl.eu/results/en&sat/corpus-result-table

[4] https://github.com/helloboyn/MMLoSo25-IT2-BT5-ES-MT

Testset	Model	BLEU (En→Sa)	chrF++ (En→Sa)	BLEU (Sa→En)	chrF++ (Sa→En)
IN22-Gen	IndicTrans2-baseline	5.5	35.8	24.8	51.0
	IndicTrans2-finetuned	**7.3**	**40.3**	**26.8**	**53.9**
	ByT5 Model	2.9	32.6	5.6	30.2
Flores200-dev	IndicTrans2-baseline	3.3	29.5	19.5	45.1
	IndicTrans2-finetuned	**4.7**	**32.7**	**22.0**	**49.2**
	ByT5-finetune	2.7	23.7	6.1	26.7

Table 4: BLEU and chrF++ evaluation scores for IndicTrans2 (baseline and finetuned) and ByT5 on the IN22-Gen and Flores200-dev test sets for both translation directions.

dings to its transformer layers, making them more powerful and efficient. The model is pre-trained with a span corruption objective that hides longer byte spans to help it understand the context better. When fine-tuning for English to Santhali translation, the encoder takes English sentences and turns them into byte sequences. The decoder then makes the Santhali translation one byte at a time. The ByT5 architecture offers several benefits, including the ability to work with any script, handle noisy or unseen inputs, and generalise effectively. This makes it a great choice for building reliable translation systems for low-resource languages, such as Santhali.

IndicTrans2 Architecture: IndicTrans2 (Gala et al., 2023) is an innovative multilingual neural machine translation (NMT) model that uses the Transformer architecture 3. A shared multilingual encoder-decoder framework enables it to work with a wide range of Indian languages. The model improves translations by leveraging the fact that Indian languages share similarities with each other. This is especially true for pairs with limited resources, such as English–Santali. The main things about IndicTrans2 are: The Transformer model is the basis for IndicTrans2. It utilises self-attention mechanisms to identify both short- and long-range dependencies within sentences. This enables the model to produce translations that are both fluent and contextually relevant. Multilingual Training: The model learns from a large amount of data that is similar across different Indian languages. This lets it learn shared representations, which helps it generalise better and makes it easier for high-resource languages to share information with low-resource languages. Shared Embeddings: IndicTrans2 utilises shared embedding spaces across languages to leverage patterns that are common to both languages. This method improves the ability to translate between languages that are significantly different from each other. Fine-Tuning Capability: The model can be fine-tuned for specific language pairs, allowing it to adjust to the unique syntax, morphology, and script of languages with limited resources, such as Santali.

5 Evaluation Metrics and Analysis

We evaluate our English to Santali and Santali to English translation models using two standard automatic metrics: BLEU (Papineni et al., 2002) and chrF++ (Popović, 2017). BLEU measures n-gram precision and is widely used for machine translation quality estimation, whereas chrF++ captures character-level similarity and is particularly effective for morphologically rich languages such as Santali.

Table 4 reports the results on two benchmark datasets, IN22-Gen with 1,024 sentences and Flores200-dev with 997 sentences. The comparison includes three systems: IndicTrans2-baseline, IndicTrans2-finetuned, and ByT5.

Across both datasets and translation directions, IndicTrans2-finetuned consistently achieves the best performance. On the IN22-Gen set, the model attains 7.3 BLEU and 40.3 chrF++ for English to Santali, and 26.8 BLEU and 53.9 chrF++ for Santali to English. In contrast, ByT5 yields 2.9 BLEU and 32.6 chrF++ for English to Santali, and 5.6 BLEU and 30.2 chrF++ for Santali to English. These results indicate that byte-level modelling is less effective in handling the Ol Chiki script and Santali morphology.

The same pattern is observed on Flores200-dev. IndicTrans2-finetuned achieves 4.7 BLEU and 32.7 chrF++ for English to Santali, and 22.0 BLEU and 49.2 chrF++ for Santali to English. ByT5 again performs considerably lower, reinforcing that subword-based multilingual pretraining is more suitable for this low-resource language pair.

Overall, the findings clearly demonstrate that IndicTrans2, especially when finetuned, provides superior translation quality for English–Santali, offering stronger lexical accuracy, improved handling of the Ol Chiki script, and better character-level consistency compared to ByT5.

6 Conclusion

In this work, we investigated English–Santali machine translation using the Ol Chiki script and conducted a focused comparison between two multilingual models: ByT5 and IndicTrans2. By isolating the English–Santali pair, we provided a clear assessment of model performance on this low-resource language. Our results show that the finetuned IndicTrans2 model consistently delivers higher BLEU and chrF++ scores than ByT5 across multiple benchmarks and translation directions. These findings highlight the advantages of subword-based multilingual pretraining for handling Santali's morphology and script-specific characteristics. Overall, our study demonstrates that careful model selection and targeted finetuning play a crucial role in improving translation quality for low-resource languages, contributing to broader efforts toward digital inclusion and linguistic preservation.

Acknowledgement

The authors express gratitude to the COIL-D Project under Bhashini, funded by MeitY, for the support and resources that facilitated the successful execution of this research.

References

2025. Multimodal models for low-resource contexts and social impact 2025. Kaggle Competition.

Ramakrishna Appicharla, Baban Gain, Santanu Pal, Asif Ekbal, and Pushpak Bhattacharyya. 2024. A case study on context-aware neural machine translation with multi-task learning. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 246–257, Sheffield, UK. European Association for Machine Translation (EAMT).

Yash Bhaskar and Parameswari Krishnamurthy. 2024. Yes-MT‘s submission to the low-resource Indic language translation shared task in WMT 2024. In *Proceedings of the Ninth Conference on Machine Translation*, pages 788–792, Miami, Florida, USA. Association for Computational Linguistics.

Raj Dabre and Anoop Kunchukuttan. 2024. Findings of WMT 2024‘s MultiIndic22MT shared task for machine translation of 22 Indian languages. In *Proceedings of the Ninth Conference on Machine Translation*, pages 669–676, Miami, Florida, USA. Association for Computational Linguistics.

Jay Gala, Pranjal A. Chitale, Raghavan AK, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar, Janki Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, Pratyush Kumar, Mitesh M. Khapra, Raj Dabre, and Anoop Kunchukuttan. 2023. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. *Preprint*, arXiv:2305.16307.

Tomasz Limisiewicz, Terra Blevins, Hila Gonen, Orevaoghene Ahia, and Luke Zettlemoyer. 2024. MYTE: Morphology-driven byte encoding for better and fairer multilingual language modeling. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15059–15076, Bangkok, Thailand. Association for Computational Linguistics.

Sebastian Nehrdich, Oliver Hellwig, and Kurt Keutzer. 2024. One model is all you need: ByT5-Sanskrit, a unified model for Sanskrit NLP tasks. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13742–13751, Miami, Florida, USA. Association for Computational Linguistics.

Partha Pakray, Santanu Pal, Advaitha Vetagiri, Reddi Krishna, Arnab Kumar Maji, Sandeep Dash, Lenin Laitonjam, Lyngdoh Sarah, and Riyanka Manna. 2024. Findings of WMT 2024 shared task on low-resource Indic languages translation. In *Proceedings of the Ninth Conference on Machine Translation*, pages 654–668, Miami, Florida, USA. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Harman Singh, Nitish Gupta, Shikhar Bharadwaj, Dinesh Tewari, and Partha Talukdar. 2024a. IndicGenBench: A multilingual benchmark to evaluate generation capabilities of LLMs on Indic languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11047–11073, Bangkok, Thailand. Association for Computational Linguistics.

Kshetrimayum Boynao Singh, Ningthoujam Avichandra Singh, Loitongbam Sanayai Meetei, Ningthoujam Justwant Singh, Thoudam Doren Singh, and Sivaji Bandyopadhyay. 2023a. A comparative study of transformer and transfer learning MT models for

English-Manipuri. In *Proceedings of the 20th International Conference on Natural Language Processing (ICON)*, pages 791–796, Goa University, Goa, India. NLP Association of India (NLPAI).

Kshetrimayum Boynao Singh, Ningthoujam Avichandra Singh, Loitongbam Sanayai Meetei, Sivaji Bandyopadhyay, and Thoudam Doren Singh. 2023b. NITS-CNLP low-resource neural machine translation systems of English-Manipuri language pair. In *Proceedings of the Eighth Conference on Machine Translation*, pages 967–971, Singapore. Association for Computational Linguistics.

Ningthoujam Justwant Singh, Kshetrimayum Boynao Singh, Ningthoujam Avichandra Singh, Sanjita Phijam, and Thoudam Doren Singh. 2024b. WMT24 system description for the MultiIndic22MT shared task on Manipuri language. In *Proceedings of the Ninth Conference on Machine Translation*, pages 797–803, Miami, Florida, USA. Association for Computational Linguistics.

Bin Wei, Zheng Jiawei, Zongyao Li, Zhanglin Wu, Jiaxin Guo, Daimeng Wei, Zhiqiang Rao, Shaojun Li, Yuanchang Luo, Hengchao Shang, Jinlong Yang, Yuhao Xie, and Hao Yang. 2024. Machine translation advancements of low-resource Indian languages by transfer learning. In *Proceedings of the Ninth Conference on Machine Translation*, pages 775–780, Miami, Florida, USA. Association for Computational Linguistics.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2022. ByT5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10:291–306.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.