Archives available at journals.mriindia.com

**International Journal on Advanced Electrical and Computer Engineering**

ISSN: 2349-9338

Volume 15 Issue 01s, 2026

# Kurukh to Hindi Translation: A Review of Methods, Challenges and Future Directions

[1]Ramesh Raj Dhruwe, [2]Sanjay Kumar, [3]Vaishali Sarde

*[1]Research Scholar, School of Studies in Computer Science and IT, Pt. Ravishankar Shukla University Raipur, India*

*[2]Professor, School of Studies in Computer Science & IT, Pt. Ravishankar Shukla University Raipur, India*

*[3]Assistant Professor, Govt. J.Yoganandam Chhattisgarh College Raipur, India*

*Email: raja.dhruwe@gmail.com, sanraipur@rediffmail.com, vaishalimirge2018@gmail.com*

**Peer Review Information**

*Submission: 05 Dec 2025*

*Revision: 25 Dec 2025*

*Acceptance: 10 Jan 2026*

**Keywords**

*Kurukh Machine Translation, Natural Language Processing, Low Resource Language Machine Translation.*

**Abstract**

Many languages have become extinct in the world today; the extinction of languages is an unfortunate challenge, as the knowledge and culture embedded in those languages also disappear. It is impossible for any human being to know all languages. The development of machine translation is a solution to overcome this limitation. Machine translation is a field of artificial intelligence and natural language processing that facilitates communication by translating from one language to another using machine translation systems. After translation, the quality of the translated output can be measured using evaluation strategies to determine whether it is of human translation quality, thus measuring its accuracy. This paper is a review of the work done on evaluating Kurukh a tribal and low-resource language machine translation challenges and future directions.

## 1. Introduction.

India is a country with a vast linguistic diversity where various cultures have flourished for a long time, with different languages and dialects spoken in different regions. Languages primarily belong to the two main families: Indo-Aryan and Dravidian. These languages are spoken by over 90% of the Indian population. According to the 2011 census, there are more than 30 languages and approximately 1369 dialects used for communication in India, along with 22 scheduled languages, including Assamese, Bengali, Bodo, Dogri, Gujarati, Hindi, Malayalam, Manipuri, Marathi, Nepali, Odia, Punjabi, Sanskrit, Kannada, Kashmiri, Konkani, Maithili, Santali, Sindhi, Tamil, Telugu, and Urdu. According to the census 2011 there are 22 Scheduled languages (Eighth Schedule) [1].

According to the methodology of the World Atlas of Languages, there are approximately 8,324 documented languages worldwide, of which around 7,000 are still actively spoken. Inspired by UNESCO's 'Atlas of the World's Languages in Danger,' Kurukh, spoken by tribal communities in India, is identified as vulnerable. Available at: https://en.wal.unesco.org (Accessed: 21 March 2025) The 2011 Census notes tribal populations comprise 8% of India's population, with a 59.0% literacy rate, matching the national average. In the state of Jharkhand, where the tribal population stands at approximately 26.21%, the Oraon community makes up 19.86%, and the Kisan community accounts for 0.43%—both of which traditionally use the Kurukh language. The number of Kurukh speakers across India is estimated to be around 1,988,350, where males 9,93,346 and females are 9,95,004 with speakers concentrated in regions such as Jharkhand, Chhattisgarh, Odisha, West Bengal, and even parts of Bangladesh. Given this context, the

present study aims to make a significant contribution to preserving and promoting Kurukh in the tribal regions (2011 Census of India).

**1.1. Origin of Kurukh:** Kurukh is a member of the Dravidian language family and is classified under the Northern Dravidian subgroup, which also includes languages such as Malto and Brahui. It is predominantly spoken by the Oraon and Kisan tribal communities in states like Jharkhand, Chhattisgarh, Odisha, West Bengal, Bihar, and Assam. Additionally, Kurukh is spoken in neighbouring countries, including Bangladesh and Nepal. Historical and linguistic research indicates that the Kurukh-speaking communities may have originally migrated from southern India and eventually settled in the central and eastern parts of the country. Although geographically distant from the core Dravidian-speaking regions, the Kurukh language has retained many of the essential linguistic characteristics typical of the Dravidian family [2]. Linguistic Features of Kurukh language: The Kurukh language has a rich phonemic system that includes a total of 20 vowel sounds comprising five short vowels, five long vowels, and their nasalized forms. In addition, it features 34 consonant sounds, which are categorized into 21 plosives, three fricatives, four nasals, four liquids, and two glides. A notable characteristic of Kurukh is its extensive use of fully phonemic aspirated consonants, which appear in both native and borrowed words. This trait is particularly uncommon among other languages in the Dravidian family [3].

**1.2. Origin of Hindi:** Hindi belongs to the Indo-Aryan branch of the Indo-European language family. Its development can be traced through stages of Classical Sanskrit, followed by Prakrit and Apabhramsha, which eventually gave rise to early forms of Hindi. The foundation of Modern Standard Hindi is primarily the Khariboli dialect, spoken in Delhi and nearby regions. Although Hindi shares much of its vocabulary and grammatical structure with Urdu, it is distinctively written in the Devanagari script. Hindi holds the status of an official language of India and is extensively used throughout North and Central India. It plays a significant role in sectors such as government, education, mass communication, and cultural life. [Online]. Available: https://www.constitutionofindia.net/articles/article-343-official-language-of-the-union/(Accessed: 15 April 2025) [4].

**1.3. Importance of translation from Kurukh to Hindi:** Translation from Kurukh to Hindi plays a vital role in promoting linguistic inclusivity, cultural preservation, and educational access for tribal communities, particularly the Oraon and Kisan tribes of eastern-central India. As Kurukh is a tribal language with limited formal recognition and resources, translating its content into Hindi, India's most widely spoken and officially recognized language, ensures broader communication and representation.
Such translation:

- Bridges communication gaps between tribal and non-tribal populations.
- Enables access to government services, education, and legal systems.
- Helps preserve tribal knowledge, folklore, and traditions in a widely understood form.
- Encourages integration without assimilation, maintaining cultural identity while enabling participation in national discourse.

**1.4. Natural Language Processing (NLP)** is a specialized area within computer science and artificial intelligence that enables machines to process and interpret human language. Its goal is to equip computers with the ability to comprehend, analyse, and generate natural language in a way that is both meaningful and useful in real-world applications [5].

**1.5. Types of Machine Translation:**

1.5.1. **Rule-Based Machine Translation (RBMT):** RBMT utilizes the predefined linguistic rules and bilingual dictionaries to perform translations Rule based MT systems works based upon the specification of rules for morphology, syntax, lexical selection and transfer and generation. Collection of rules and a bilingual or multilingual lexicon are the resources used in RBMT. In the case of English to Indian languages and Indian language to Indian language MT systems, there have been many attempts with all these approaches [6]. The transfer model involves three stages: analysis, transfer and generation. Figure 1 shows the complete workflow of translation in the form of a pipeline. During analysis phase linguistic analysis is performed on the input source sentence in order to extract information in terms of morphology, parts of speech, phrases, named entity and word sense disambiguation. During the lexical transfer phase, there are two steps namely word translation and grammar translation. In word translation, source language root word is replaced by the target language root word with the help of a bilingual dictionary and in grammar

translation, suffixes are getting translated. In generation phase genders of the translated words are corrected and it will be followed by short distance and long-distance agreements performed by intra-chunk and the inter-chunk module. These ensure that the gender, number and person of local groups of phrases agree as also the gender of the subject's verbs or objects reflect those of the subject.[7]

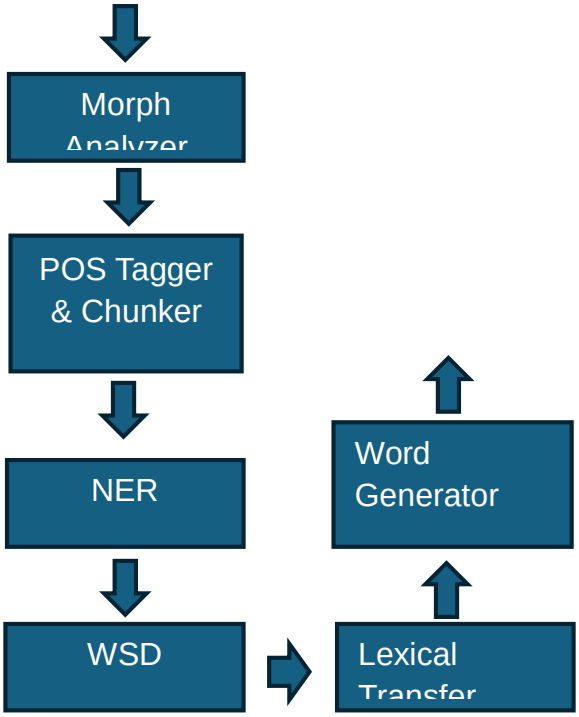


*Fig 1: RBMT Workflow*

1.5.2. **Statistical Machine Translation (SMT):** SMT relies on statistical models derived from large bilingual corpora [8]. The statistical approach works based up on the statistical models extracted from parallel aligned bilingual text corpora, which takes the assumption that every word in the target language is a translation of the source language words with some probability The words which have the highest probability will give the best translation. Consistent patterns of divergence between the languages, when translating from one language to another, handling reordering divergence are one of the fundamental problems in MT. Figure 2 shows the functional flow diagram of an SMT system. The major steps in SMT are: Corpus preparation, Training. Decoding and Testing. Corpus preparation, alignment and its cleaning will be done in the Pre-Processing step. Training is a process in which a supervised or unsupervised statistical machine learning algorithm is used to build statistical tables from the parallel corpora. In Statistical, Machine Translation, word by word and phrase based alignment plays the major role during parallel corpus training. During training Translational model, Language Model, Distortion Table, Phrase table etcetera are modelled. Decoding is the most complex task in Machine Translation where the trained models will be decoded. It is the major process in which the target language translations are being decoded using the generated phrase table, translation model and language model. The two major concerns with SMT are decoding complexity and target language reordering [7].

1.5.3. **Neural Machine Translation (NMT):** NMT employs deep learning techniques, particularly neural networks, to model the entire translation process Neural Machine Translation (NMT) is an advanced approach to Machine Translation (MT) that uses deep neural networks to automatically translate text from one language to another. Unlike earlier rule-based and statistical methods, NMT works as an end-to-end system without the need for handcrafted linguistic rules or separate language components. Its architecture is primarily built on the encoder–decoder model with recurrent neural networks such as Long Short-Term Memory (LSTM). The encoder converts the input sentence into a contextual representation, while the decoder generates the target translation using attention mechanisms that allow the model to focus on relevant parts of the source sentence

during translation. Pathak and Pakray (2019) examined NMT performance for Indian languages by training systems for English-to-Hindi, English-to-Tamil, and English-to-Punjabi using OpenNMT. Their human and BLEU evaluation results indicated that NMT produces more fluent translations, especially for Hindi and Punjabi. However, performance was weaker for Tamil due to its agglutinative and morphologically complex structure, which increases the number of unknown words during translation. The study also showed that translation quality improves with larger training data and longer sentences, due to richer contextual learning. Overall, the work highlights NMT's potential for low-resource Indian languages while emphasizing challenges related to linguistic diversity.[9] [10].

**Diagram: NLP Workflow: -**

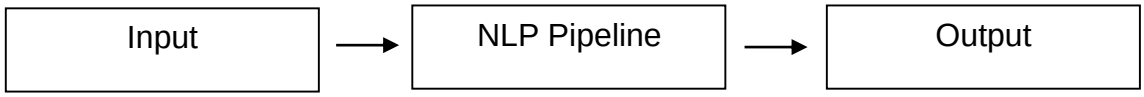


*Fig 2: NLP Workflow of Translation System.*

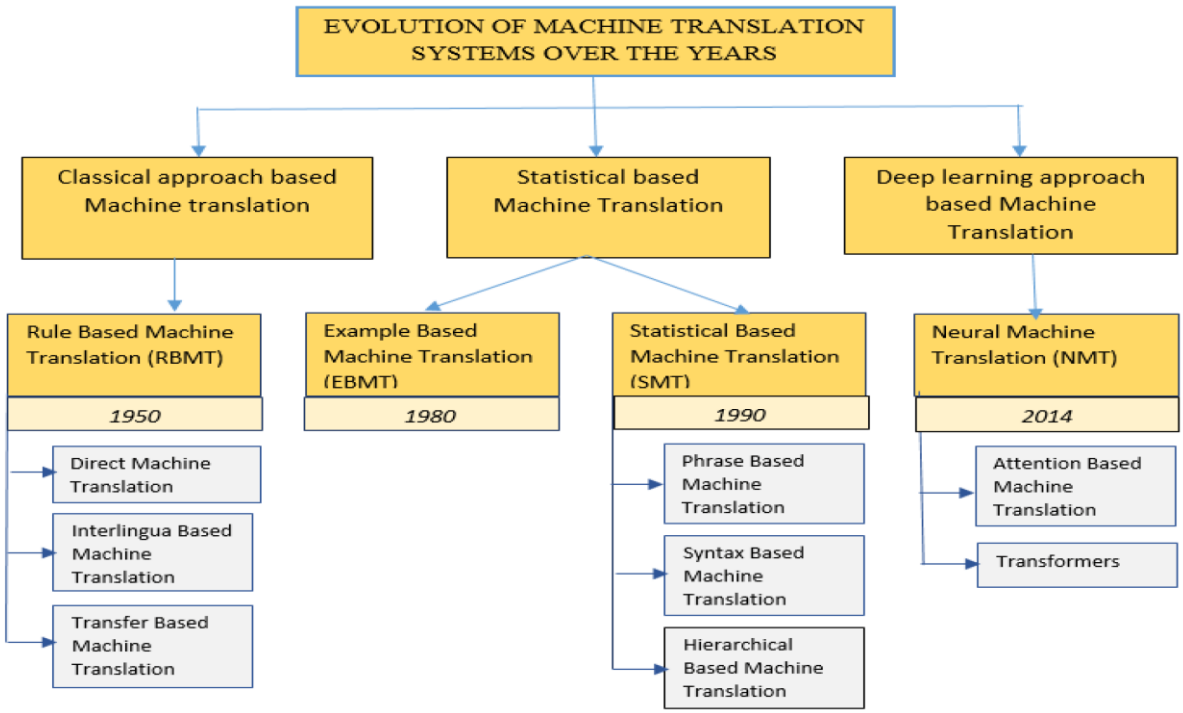


*Fig 3. The taxonomy is a representation of the evolving machine translation systems over the year [11].*

1.5.4. **NMT (Neural Machine Translation) Overview:**

NMT is a modern technique that uses deep learning to translate text between languages. Unlike older methods that break the process into separate steps, NMT uses a single, end-to-end neural network usually built with encoder-decoder models and attention mechanisms. This setup helps the system understand the meaning and context of entire sentences, not just individual words or phrases. As a result, translations are smoother and more accurate. Over the years, NMT has become the go-to approach in both research and real-world applications because of its high performance and natural-sounding output [12].

Transformer-Based NMT (Modern Approach): Instead of RNNs, transformers (like NLLB-200 [13]) use attention mechanisms.

**Diagram: Transformer Model for NMT**

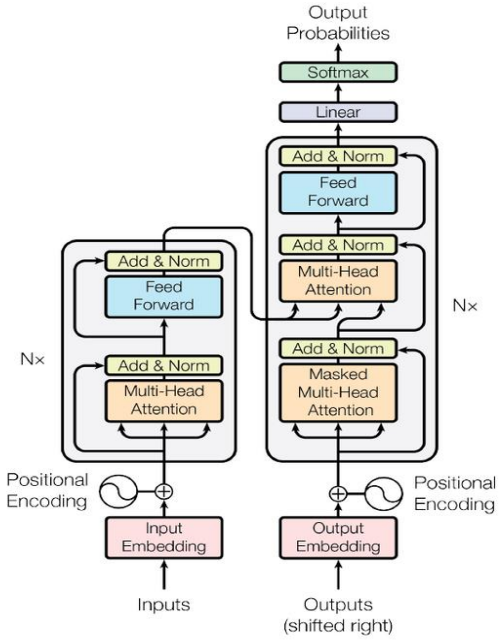


Fig 4: Transformer Model

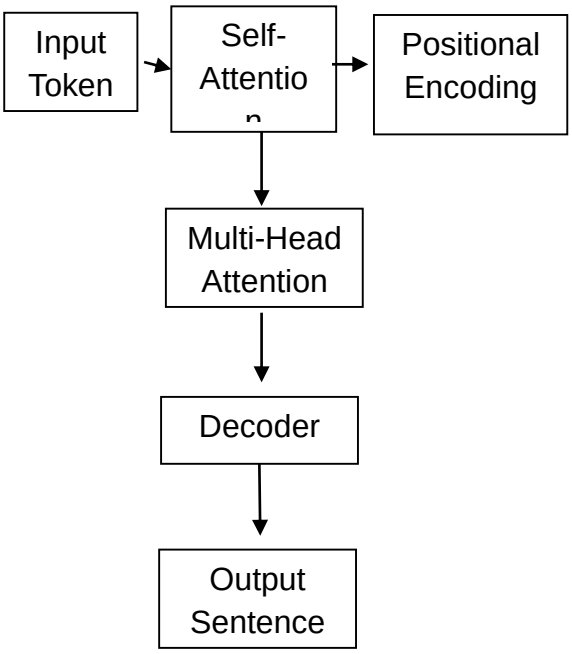


*Fig 5: Transformer Model for NMT*

layer normalization. Its simplicity, efficiency, and scalability have since made it a foundation for many modern NLP models [14].

## 2. Literature Review.

A novel approach to machine translation (MT) by combining the strengths of Neural Machine Translation (NMT) and Statistical Machine Translation (SMT). The authors propose a deep neural network-based system combination framework that leverages minimum Bayes-risk decoding and multi-source NMT to integrate the N-best outputs of both NMT and SMT systems. This approach is applied to both RNN and self-attention (Transformer) models, with experiments conducted on resource-rich Chinese-English and low-resource English-Vietnamese translation tasks with the proposed model achieving higher BLEU scores and better handling of rare and unknown words [15].

The application of neural machine translation (NMT) for Indian languages, comparing its effectiveness against statistical machine translation (SMT), highlight the challenges posed by the linguistic diversity of Indian languages, particularly the differences between Indo-Aryan and Dravidian language families. It evaluates various sub word segmentation techniques, with a focus on Byte Pair Encoding (BPE), and finds that lower BPE merge operations (0–5000) yield optimal results for Indian languages, unlike European languages, where higher merge values are preferred. Contributing valuable insights into the development of efficient NMT systems for low-resource languages. The findings suggest that hybrid approaches and language-specific adaptations can further enhance translation accuracy [16].

Statistical Machine Translation (SMT) system for Assamese-English using Moses, IRSTLM, and GIZA++. With a BLEU score of 9.72 (English-Assamese: 5.02), it includes a transliteration module for OOV words. Future work focuses on corpus expansion, improved transliteration, parallel processing and a web-based user interface [17],[18].

A translation system based on the Anglabharti methodology. AnglaHindi is a hybrid system that combines rule-based and example-based approaches, integrating statistical methods to enhance translation accuracy from English to Hindi. It follows a pseudo-interlingual approach. The system leverages multilingual dictionaries, syntactic and semantic rules, and a correction module to improve translation quality. The system, available online, achieves approximately 90% accuracy for simple and complex sentences up to 20 words. The importance of hybrid methodologies in overcoming linguistic challenges for English and Hindi [19].

Anusaaraka, a hybrid MT system based on Paninian Grammar, prioritizing translation accuracy over fluency. It integrates Apertium with rule-based linguistic processing and allows user contributions for word sense disambiguation. The future includes debugging tools and gamified WSD rule creation, emphasizing collaborative, knowledge-driven translation for Indian languages [20].

A clear, comprehensive introduction to machine translation (MT) for non-specialists, including translators and students. It explores MT history, system architectures, translation challenges, and evaluation methods. With structured explanations, summaries, and bibliographies, it supports deeper learning, though coverage of AI-based approaches could be more detailed [21].

Statistical Machine Translation (SMT) model for English to South Dravidian languages like Malayalam and Kannada. Addressing syntactic and morphological challenges, it uses reordering, root-suffix separation, and morphological features to enhance translation accuracy. Employing EM algorithms, IBM models, SRILM, and MOSES, the approach improves BLEU scores and is extendable to other Dravidian languages, aiding low-resource translation research [22].

How NLP impacts information retrieval (IR), concluding that statistical "bag of words" approaches outperform deep linguistic methods like syntactic and semantic analysis. Linguistic enhancements often degrade performance due to weighting complexities. NLP proves more beneficial in tasks like question answering and summarization. The paper also covers IR evaluation using test collections, precision, and recall metrics [23].

Reviews machine translation (MT) methods for Indian languages, including direct, transfer-based, interlingua, statistical, and example-based approaches. Projects like Anusaaraka, Anglabharti, and Shakti are highlighted. It identifies challenges such as morphology, syntax variation, and resource scarcity. The study concludes hybrid systems combining rule-based and statistical methods are most effective for India's multilingual landscape [24].

An automatic post-editing (APE) approach to improve rule-based MT using statistical models. Testing on English–Spanish corpora show APE enhances fluency, reducing editing effort. Parliament corpus performance improved by 59.5%, and Protocols by 6.5%. The approach combines RBMT's robustness with SMT refinement, offering promise for adaptive translation systems and hybrid MT development [25].

A large-scale English–Vietnamese parallel corpus to support MT systems is build. It includes over 800,000 sentence pairs from books, news, and legal texts, annotated with POS tags and word alignment. Using a semi-automated tool, BiCA2012, the corpus boosts MT accuracy for resource-poor Vietnamese. Future work will enhance alignment and integrate semantic and syntactic features [26].

HINDIA, a deep learning-based Hindi spell-checker using BiRNN and LSTM. Unlike rule-based models, HINDIA handles real-world errors by understanding context, achieving 80% accuracy. Trained on an IIT Bombay corpus, it outperforms existing models. The system illustrates deep learning's effectiveness for morphologically rich languages like Hindi in NLP applications [27].

SMT for Sinhala–Tamil translation, showing better results than English–Sinhala systems due to shared linguistic structures. Built using

GIZA++ and IBM models, the system benefits from a bilingual corpus. The research confirms SMT's advantage for related languages and suggests that larger parallel corpora can further boost translation performance for South Asian languages [28].

Ferdinand Hahn made foundational contributions to Kurukh grammar by documenting its phonetics, morphology, and syntax. His detailed analysis includes verb classifications, noun declensions, and sentence structure, offering critical insight into this Dravidian language spoken by the Oraon tribe. Hahn's work has been instrumental in preserving Kurukh and serves as a key reference for future linguistic studies and educational development [29].

A rule-based MT system for Chhattisgarhi-to-Hindi translation to support communication in Chhattisgarh is build. It includes a dictionary, POS tagger, morphological analyser, and parser using 500 rules and 30 tag categories. Achieving 80% accuracy supports domains like literature and news. The system emphasizes preserving Chhattisgarhi and demonstrates effective low-resource translation through structured linguistic methods [30].

English-to-Dravidian language translation using a sequence-to-sequence stacked LSTM model. They address morphological complexity and agglutination in Dravidian languages. Extensive dataset cleaning and experimentation with hyperparameters led to BLEU score-based evaluation. The Tamil-Telugu model ranked second with a score of 0.43. The study emphasizes the challenges in translating low-resource languages and recommends further NMT advancements for accuracy and efficiency [31].

An overview of machine translation systems and their evaluation, tracking the shift from SMT to NMT. They highlight improvements in translation accuracy and error reduction with neural methods. The paper identifies challenges such as low-resource languages and underlines the importance of evaluation metrics to assess translation quality. It concludes that advanced models can significantly enhance multilingual communication [32].

The challenge of translating languages with limited parallel data. Techniques include leveraging monolingual data via back-translation, joint training, and unsupervised methods, as well as exploiting auxiliary languages through multilingual training and transfer learning. Multi-modal data, like images or speech, also aids translation. Despite progress, challenges remain, such as selecting optimal auxiliary languages, handling unseen vocabularies, and improving pivot-based methods. Future directions include better integration of bilingual dictionaries and speech data. This survey highlights innovative approaches to enhance NMT for low-resource languages, offering insights for researchers and practitioners [33].

Transformer-based NMT for four Asian languages (Japanese, Lao, Malay, Vietnamese) with extremely limited parallel data (<20k sentences). Tagged back-translation using monolingual data outperformed other methods, achieving significant BLEU gains. Hyperparameter tuning proved critical. Results highlight challenges in low-resource settings but offer practical strategies for improvement [34].

A Universal Neural Machine Translation (NMT) approach for low-resource languages, leveraging transfer learning to share lexical and sentence-level representations across languages. It employs Universal Lexical Representation (ULR) for word-level sharing and a Mixture of Language Experts (MoLE) for sentence-level sharing. Evaluated on Romanian-English with only 6k parallel sentences, it achieves 23 BLEU, outperforming baselines. The method also supports zero-shot translation and fine-tuning, demonstrating robust performance with minimal data [35].

A neural machine translation (NMT) model for low-resource languages by incorporating local dependencies and word alignments is purposed. Their approach outperforms baseline NMT with limited data (70k tokens) but remains behind SMT in performance. The study highlights challenges in low-resource NMT [36].

**Table 1:** The overall POS Tagging review is outlined

| Reference | MT / NLP Type | Model Used | Language(s) | Accuracy / BLEU Score |
|---|---|---|---|---|
| Shrivastava et al. 2008, [40] | POS Tagging | HMM + Naïve Stemmer | Hindi | 93.12% Accuracy |

| | | | | |
|---|---|---|---|---|
| Dinesh Kumar et al. 2010, [36] | POS Tagging Review | HMM, CRF, SVM | Multiple Indian Languages | Accuracy depends on corpus, model |
| Joshi et al. 2013, [38] | POS Tagging | Hidden Markov Model (HMM) | Hindi | 92% Accuracy |
| Kumawat et al. 2015 [39] | POS Tagging | HMM, Trigram | Hindi | 95.45% Accuracy |

**Table 2:** The overall Machine Translation is outlined

| **Reference** | **MT / NLP Type** | **Model Used** | **Language(s)** | **Accuracy / BLEU Score** |
|---|---|---|---|---|
| Ferdinand Hahn et al. 1905, [29] | Manuscript (Grammar Documentation) | Phonetics, Morphology, Syntax | Kurukh-Hindi | - |
| Hegde et al. 2021, [37] | NMT | LSTM | Tamil-Telugu | BLEU: 0.43 |
| Pandey Vikas et al. 2023, [30] | RBMT | Rule-Based (500 rules, POS tagger, Morph Analyzer) | Chhattisgarhi-Hindi | 80% Accuracy |
| Long Zhou et al. 2003, [6] | NMT + SMT (Hybrid) | Deep Neural Networks, MBR decoding | English-Vietnamese | Improved BLEU (score not stated) |
| Kalyanee Kanchan Baruah et al. 2014, [38] | SMT | Moses toolkit, IRSTLM, GIZA++ | English-Assamese | BLEU: 9.72 |
| Joshi et al. 2013, [39] | POS Tagging | Hidden Markov Model (HMM) | Hindi | 92% Accuracy |
| Sinha et al. 1995, [19] | Hybrid MT | AnglaHindi (Rule + Example-Based) | English-Hindi | ~90% Accuracy (≤ 20 words) |
| Chaudhury et al. 2010, [20] | Hybrid MT | Anusaaraka (Paninian Grammar + Apertium) | Hindi, Bengali, Telugu, Kannada | Accuracy-focused (BLEU not stated) |
| Dewangan et al. 2021, [16] | NMT vs. SMT | Transformer, BPE Segmentation | Hindi, Marathi, Tamil, Telugu | BLEU improved (lower BPE optimal) |
| Ruvan Weerasinghe et al. 2010 [28] | SMT | IBM Models, GIZA++ | Sinhala–Tamil | Outperforms English–Sinhala (BLEU not stated) |

3. **Challenges for Machine Translation:**

**Data Scarcity:** There is a lack of substantial parallel corpora and linguistic resources for training effective NMT model for Kurukh-to-Hindi translation. The development of high-quality datasets, including domain-specific texts and conversational data, is crucial for improving translation accuracy and fluency.

**Script Adaptation:** Kurukh traditionally uses the Devanagari script, but the use of the Devanagari script introduces complexities in script standardization and character mapping. Limited research is available on how to effectively adapt Devanagari script resources to accurately represent Kurukh's linguistic features, phonology, and grammar.

**Interlinguistic Challenges:** The linguistic structures of Hindi and Kurukh differ significantly, especially in terms of syntax, morphology, and semantics. The current National Linguistic Standard (NMT) model often struggles with these differences, and innovative approaches are needed to effectively model cross-linguistic variations and semantic nuances.

**Evaluation Criteria:** Standard evaluation criteria for translation quality, such as the BLEU score, may not fully reflect the quality of translations in low-resource languages like Kurukh. Developing new evaluation frameworks that better assess the accuracy, fluency, and cultural relevance of translations is an area that needs further exploration.

## 4. Conclusion.

While significant advancements have been made in neural machine translation (NMT) systems for widely spoken language pairs, such as English to Hindi or Hindi to Bengali, there remains a notable research gap in the domain of NMT for less-resourced languages, particularly for the translation from Kurukh to Hindi using the Devanagari script. Kurukh, a Dravidian language spoken by the Oraon people primarily in parts of India, is underrepresented in the NMT landscape due to its limited digital resources and linguistic resources. Moreover, the use of Devanagari script for Kurukh presents unique challenges in script adaptation, lexical alignment, and syntactic divergence that have not been adequately addressed in existing studies. Due to data constraints: There is a lack of sufficient parallel corpora and linguistic resources for training effective NMT models for Kurukh to Hindi translation. Corpora can be developed from various sources of linguistic resources, such as books, magazines, poetry, newspapers, story collections, etc. Developing high-quality datasets, including domain-specific text and conversational data, is crucial for improving translation accuracy and fluency.

## References.

[1] S. Khanuja *et al.*, “MuRIL: Multilingual Representations for Indian Languages,” Apr. 02, 2021, *arXiv*: arXiv:2103.10730. doi: 10.48550/arXiv.2103.10730.

[2] K. Bhadriraju, *The Dravidian Languages*. Cambridge University Press, 2003.

[3] B. P. Mahapatra, *The Kurux Language: A Descriptive Study*. Mysore, India: Central Institute of Indian Languages, 1989.

[4] C. P. Masica, *The Indo-Aryan languages*. Cambridge ; New York : Cambridge University Press, 1931. [Online]. Available: https://archive.org/details/indoaryanlanguag0000masi

[5] Kavuri, S. (2023). Machine learning approaches for security vulnerability detection in software testing. Computer Fraud & Security, 2023(10). https://doi.org/10.52710/cfs.837

[6] S. Dave, J. Parikh, and P. Bhattacharyya, “Interlingua-based English–Hindi Machine Translation and Language Divergence”.

[7] S. Sreelekha, P. Bhattacharyya, and D. Malathi, “Statistical vs. Rule-Based Machine Translation: A Comparative Study on Indian Languages,” in *International Conference on Intelligent Computing and Applications*, vol. 632, S. S. Dash, S. Das, and B. K. Panigrahi, Eds., in Advances in Intelligent Systems and Computing, vol. 632. , Singapore: Springer Singapore, 2018, pp. 663–675. doi: 10.1007/978-981-10-5520-1_59.

[8] P. Koehn, F. J. Och, and D. Marcu, “Statistical Phrase-Based Translation,” *Proceedings of HLT-NAACL 2003*, pp. 48–54, June 2003.

[9] A. Pathak and P. Pakray, “Neural Machine Translation for Indian Languages,” *Journal of Intelligent Systems*, vol. 28, no. 3, pp. 465–477, July 2019, doi: 10.1515/jisys-2018-0065.

[10] D. Bahdanau, K. Cho, and Y. Bengio, “Neural Machine Translation by Jointly Learning to Align and Translate,” May 19, 2016, *arXiv*: arXiv:1409.0473. doi: 10.48550/arXiv.1409.0473.

[11] S. Sharma, M. Diwakar, P. Singh, V. Singh, S. Kadry, and J. Kim, “Machine Translation Systems Based on Classical-Statistical-Deep-Learning Approaches,” *Electronics*, vol. 12, no. 7, p. 1716, Apr. 2023, doi: 10.3390/electronics12071716.

[12] Z. Tan *et al.*, “Neural machine translation: A review of methods, resources, and tools,” *AI Open*, vol. 1, pp. 5–21, 2020, doi: 10.1016/j.aiopen.2020.11.001.

[13] Y. Koishekenov, A. Berard, and V. Nikoulina, “Memory-efficient NLLB-200: Language-specific Expert Pruning of a Massively Multilingual

Machine Translation Model," July 07, 2023, *arXiv*: arXiv:2212.09811. doi: 10.48550/arXiv.2212.09811.
[14] A. Vaswani *et al.*, "Attention Is All You Need," Aug. 02, 2023, *arXiv*: arXiv:1706.03762. doi: 10.48550/arXiv.1706.03762.
[15] L. Zhou, J. Zhang, X. Kang, and C. Zong, "Deep Neural Network--based Machine Translation System Combination," *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, vol. 19, no. 5, pp. 1–19, Sept. 2020, doi: 10.1145/3389791.
[16] S. Dewangan, S. Alva, N. Joshi, and P. Bhattacharyya, "Experience of neural machine translation between Indian languages," *Machine Translation*, vol. 35, no. 1, pp. 71–99, Apr. 2021, doi: 10.1007/s10590-021-09263-3.
[17] K. Kanchan Baruah, P. Das, A. Hannan, and S. Kr Sarma, "Assamese-English Bilingual Machine Translation," *IJNLC*, vol. 3, no. 3, pp. 73–82, June 2014, doi: 10.5121/ijnlc.2014.3307.
[18] S. Kumar, "Mathematical Modelling and Simulation of Fault Tolerant Double Tree Network," presented at the International Conference on Advanced Computing and Communications, pp. 422–431. doi: 10.11.1109/ADCOM.2007.62.
[19] R. M. K. Sinha, K. Sivaraman, A. Agrawal, R. Jain, R. Srivastava, and A. Jain, "ANGLABHARTI: a multilingual machine aided translation project on translation from English to Indian languages," in *1995 IEEE International Conference on Systems, Man and Cybernetics. Intelligent Systems for the 21st Century*, Vancouver, BC, Canada: IEEE, 1995, pp. 1609–1614. doi: 10.1109/ICSMC.1995.538002.
[20] S. Chaudhury, A. Rao, and D. M. Sharma, "Anusaaraka: An expert system based machine translation system," in *Proceedings of the 6th International Conference on Natural Language Processing and Knowledge Engineering(NLPKE-2010)*, Beijing, China: IEEE, Aug. 2010, pp. 1–6. doi: 10.1109/NLPKE.2010.5587789.
[21] D. J. Arnold and D. Arnold, Eds., *Machine translation: an introductory guide*, 1. publ. Manchester: NCC Blackwell, 1994.
[22] A. V. Vidyapeetham, "A Novel Approach for English to South Dravidian Language Statistical Machine Translation System," vol. 02, no. 08, 2010.
[23] E. M. Voorhees, "Natural Language Processing and Information Retrieval," in *Information Extraction*, vol. 1714, M. T. Pazienza, Ed., in Lecture Notes in Computer Science, vol. 1714. , Berlin, Heidelberg: Springer Berlin Heidelberg, 1999, pp. 32–48. doi: 10.1007/3-540-48089-7_3.
[24] L. R.Nair and D. Peter S., "Machine Translation Systems for Indian Languages," *IJCA*, vol. 39, no. 1, pp. 24–31, Feb. 2012, doi: 10.5120/4785-7014.
[25] A.-L. Lagarda, V. Alabau, F. Casacuberta, R. Silva, and E. Díaz-de-Liaño, "Statistical post-editing of a rule-based machine translation system," in *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers on - NAACL '09*, Boulder, Colorado: Association for Computational Linguistics, 2009, p. 217. doi: 10.3115/1620853.1620913.
[26] Q. H. Ngo and W. Winiwarter, "Building an English-Vietnamese Bilingual Corpus for Machine Translation," in *2012 International Conference on Asian Language Processing*, Hanoi, Vietnam: IEEE, Nov. 2012, pp. 157–160. doi: 10.1109/IALP.2012.30.
[27] S. Singh and S. Singh, "HINDIA: a deep-learning-based model for spell-checking of Hindi language," *Neural Comput & Applic*, vol. 33, no. 8, pp. 3825–3840, Apr. 2021, doi: 10.1007/s00521-020-05207-9.
[28] S. Sripirakas, A. R. Weerasinghe, and D. L. Herath, "Statistical machine translation of systems for Sinhala - Tamil," in *2010 International Conference on Advances in ICT for Emerging Regions (ICTer)*, Colombo, Sri Lanka: IEEE, Sept. 2010, pp. 62–68. doi: 10.1109/ICTER.2010.5643268.
[29] T. Revd. FERD. HAHN, *Kurukh Grammar*, 2nd ed. Calcutta: Bengal Secretariat Press, 1911.
[30] V. Pandey, "Computer Science & Engineering Faculty of Computer & Information Technology," *Computer Science*, 2023.
[31] A. Hegde, I. Gashaw, and H. L. Shashirekha, "MUCS@ - Machine Translation for Dravidian Languages using Stacked Long Short Term Memory," *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, p. pages 340-345, Apr. 2021.
[32] S. Singh and S. Singh, "HINDIA: a deep-learning-based model for spell-checking of Hindi language," *Neural Comput & Applic*, vol. 33, no. 8, pp. 3825–3840, Apr. 2021, doi: 10.1007/s00521-020-05207-9.
[33] R. Wang, X. Tan, R. Luo, T. Qin, and T.-Y. Liu, "A Survey on Low-Resource Neural Machine Translation," July 09, 2021, *arXiv*: arXiv:2107.04239. doi: 10.48550/arXiv.2107.04239.
[34] R. Rubino, B. Marie, R. Dabre, A. Fujita, M. Utiyama, and E. Sumita, "Extremely low-resource neural machine translation for Asian languages," *Machine Translation*, vol. 34, no. 4, pp. 347–382, Dec. 2020, doi: 10.1007/s10590-020-09258-6.
[35] J. Gu, H. Hassan, J. Devlin, and V. O. K. Li, "Universal Neural Machine Translation for Extremely Low Resource Languages," Apr. 17,

2018, *arXiv*: arXiv:1802.05368. doi: 10.48550/arXiv.1802.05368.
[36] R. Östling and J. Tiedemann, "Neural machine translation for low-resource languages," Aug. 18, 2017, *arXiv*: arXiv:1708.05729. doi: 10.48550/arXiv.1708.05729.
[37] A. Hegde, I. Gashaw, and H. L. Shashirekha, "MUCS@ - Machine Translation for Dravidian Languages using Stacked Long Short Term Memory," p. pages 340-345, Apr. 2021.
[38] K. Kanchan Baruah, P. Das, A. Hannan, and S. Kr Sarma, "Assamese-English Bilingual Machine Translation," *IJNLC*, vol. 3, no. 3, pp. 73–82, June 2014, doi: 10.5121/ijnlc.2014.3307.
[39] N. Joshi, H. Darbari, and I. Mathur, "HMM Based POS Tagger for Hindi," in *Computer Science & Information Technology ( CS & IT )*, Academy & Industry Research Collaboration Center (AIRCC), Sept. 2013, pp. 341–349. doi: 10.5121/csit.2013.3639.