

English–Nepali–Tamang: A Trilingual Parallel Corpus and Benchmark for Low-Resource Machine Translation

Praveen Acharya
Dublin City University
acharyaprvn@gmail.com

Rupak Raj Ghimire
Kathmandu University
rughimire@gmail.com

Prakash Poudyal
Kathmandu University
prakash@ku.edu.np

Balaram Prasain
Tribhuvan University
prasainbalaram@gmail.com

Bal Krishna Bal
Kathmandu University
bal@ku.edu.np

Abstract

This article describes the research project aimed at developing a Trilingual Machine Translation System for English, Nepali, and Tamang language pairs.

1 Introduction

The predominance of English as the primary language of digital content creates a significant barrier for speakers of Nepali and minority languages such as Tamang. This linguistic divide restricts equitable access to information, knowledge resources, and digital services across sectors, including education, governance, healthcare, and commerce. Machine Translation (MT) provides a scalable solution to bridge this gap. However, MT systems for low-resource languages remain under-developed due to the scarcity of high-quality parallel corpora. Research on MT for English–Nepali–Tamang language pairs has been limited. Early work on English–Nepali MT, including rule-based (Bista et al., 2005) and lexicon-driven approaches (Dahal, 2011), showed limited scalability. Subsequent studies comparing Statistical Machine Translation (SMT) and Neural Machine Translation (NMT) found that SMT performs better for English–Nepali, mainly due to the lack of sufficient parallel data for training NMT models (Acharya and Bal, 2018). While some domain-specific datasets exist for English–Nepali (Poudel et al., 2024), resources for Nepali–Tamang remain scarce. A small parallel corpus of about 15,000 sentence pairs is available (Chaudhary et al., 2020), and there is currently no parallel corpus for the English–Tamang language pair. This

project addresses these challenges by developing a community-driven trilingual parallel corpus (Bal et al., 2024) and an extensible MT system for English, Nepali, and Tamang, with a strong focus on inclusivity, scalability, and real-world impact.

1.1 Objectives

The key objectives of this research project are:

- **Corpus Development Framework:** Design and implement a sustainable, community-driven mechanism for collecting and validating parallel corpora for low-resource languages.
- **Community Engagement:** Ensure active participation of native speakers, translators, and local stakeholders throughout to promote ownership and inclusivity.
- **Extensible MT System:** Develop a scalable multilingual MT system that is adaptable to additional languages in Nepal.

1.2 Funding Details

- **Project Title:** Empowering Information Access Rights: Developing a Trilingual Machine Translation System for English, Nepali, and Tamang.
- **Funding Agency:** Google Academic Research Award (GARA) Society-Centered AI
- **Award Year:** 2024 A.D.
- **Partner Institutions:** Kathmandu University and Tribhuvan University
- **Principal Investigator:** Bal Krishna Bal
- **Co-Principal Investigator:** Balaram Prasain

- **Contributors:** Bipesh Subedi, Jenny Poudel, Nischal Karki, Prakash Poudyal, Praveen Acharya, Rishikesh Kumar Sharma, Rupak Raj Ghimire, Rupak Tiwari.

2 Expected Outcomes

The project is expected to deliver the following:

- **Parallel Corpora:** Approximately 100,000 sentence pairs for each language pair.
- **MT System Framework:** A scalable and extensible multilingual translation system.
- **Community Awareness:** Increased awareness of language technologies for preservation and digital inclusion.
- **Adoption and Impact:** Enhanced use in local and provincial governments, community schools, and public service delivery.
- **Language Preservation:** Strengthening the digital presence of Tamang and other under-represented languages.

3 Progress and Current Outcomes

MT System: A beta version¹ of the system has been released for community testing.

Dataset and MT models: A large-scale Nepali–Tamang parallel dataset: (1) *NepTam20K*, a corpus of 20,000 sentence pairs translated by experts, and (2) *NepTam80K*, a synthetic corpus of 80,000 sentence pairs.

Both datasets are used to fine-tune state-of-the-art multilingual MT models (mBART, M2M-100, and NLLB-200). Among the evaluated models, the fine-tuned NLLB-200 model outperformed all other models across metrics in both translation directions for the Tamang–Nepali language pair. Details of the corpus creation pipeline, data composition, methodology, and experimental evaluation are presented in (Ghimire et al., 2026).²

4 Next Steps

The next phase of the project will focus on:

- Extending the corpus to the target of 100,000 sentence pairs by June 2026.

- Enhancing MT system performance through iterative refinement.
- Promoting adoption through collaboration with local and provincial governments.
- Increasing community engagement to ensure sustained usage and impact.

Acknowledgement: This research was funded by Google through the 2024 Google Academic Research Award (GARA) under the Society-Centered AI initiative and Taighde Éireann – Research Ireland under Grant No. 18/CRT/6223.

References

- Acharya, Praveen and Bal Krishna Bal. 2018. A Comparative Study of SMT and NMT: Case Study of English–Nepali Language Pair. In *6th Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU)*, pages 90–93.
- Bal, Bal Krishna, Balaram Prasain, Rupak Raj Ghimire, and Praveen Acharya. 2024. Strategies for corpus development for low-resource languages: Insights from nepal. *Automatic Speech Recognition and Translation for Low Resource Languages*, pages 297–330.
- Bista, S, B Keshari, J Bhatta, and K Parajuli. 2005. Dobhase: online English to Nepali machine translation system. In *The proceedings of the 26th Annual conference of the Linguistic Society of Nepal*.
- Chaudhary, Binaya Kumar, Bal Krishna Bal, and Rasil Baidar. 2020. Efforts towards developing a Tamang Nepali machine translation system. In *Proceedings of the 17th international conference on natural language processing (ICON)*, pages 281–286.
- Dahal, AR. 2011. Development of a Nepali–English MT system using the apertium MT platform. *The Language Technology Kendra*.
- Ghimire, Rupak Raj, Bipesh Subedi, Balaram Prasain, Prakash Poudyal, Praveen Acharya, Nischal Karki, Rupak Tiwari, Rishikesh Kumar Sharma, Jenny Poudel, and Bal Krishna Bal. 2026. Nep-Tam: A Nepali–Tamang Parallel Corpus and Baseline Machine Translation Experiments. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 8850–8861, Palma, Mallorca, Spain, May. European Language Resources Association (ELRA).
- Poudel, Shabdapurush, Bal Krishna Bal, and Praveen Acharya. 2024. Bidirectional English–Nepali machine translation (MT) system for legal domain. In *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages@ LREC-COLING 2024*, pages 53–58.

¹<https://tmt.ilprl.ku.edu.np/>

²<https://github.com/ilprl/NepTam-A-Nepali-Tamang-Parallel-Corpus-and-Baseline-Machine-Translation-Experiments>