

Parallel Corpus Development Toolkit (PCDT): A Web-Based Platform for Multilingual Parallel Data Creation

Praveen Acharya
Dublin City University
acharyapravn@gmail.com

Rupak Raj Ghimire
Kathmandu University
rughimire@gmail.com

Bipesh Subedi
Kathmandu University
bipeshrajsubedi@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 paper presents *PCDT*, a web-based platform for collecting sentence-aligned parallel corpora through a community-driven approach to support machine translation for under-resourced languages. The tool decentralizes the translation task to the target community and is subsequently reviewed by language experts.

1 Introduction

The purpose of the Parallel Corpus Development Toolkit (*PCDT*) is to create a web-based platform that streamlines the development of multilingual parallel corpora for machine translation systems, with Nepali as the primary language. The platform enables *administrators* to manage tasks and text corpora, *translators* to provide multilingual translations, and *reviewers* to verify translation quality, while also supporting the addition of new languages for future scalability. There are few systems (Zanata,¹ TEITOK,² Pootle,³ WebAnno (Yimam et al., 2013)) popular for the translation task. These are mostly designed for application localization and linguistic annotation, and hence are not necessarily suitable for sentence-aligned corpus translation. Furthermore, these systems lack the multiversioning of the edits and recoverability, which is essential for mass deployment.

The scope of the application is to serve as a centralized platform for parallel corpus development in different languages, facilitating the creation of

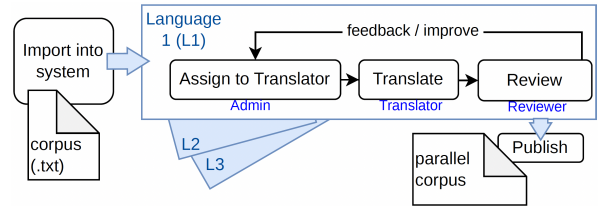


Figure 1: Workflow used in Parallel Corpus Development Toolkit.

high-quality multilingual parallel corpora required for training a multilingual Neural Machine Translation (NMT) system, including Nepali, English, and Tamang, with potential expansion to other languages. The objective is to support community-driven parallel corpus development, a promising strategy for developing corpora in low-resource languages (Bal et al., 2024).

2 System Overview

PCDT is a web-based application with three main user roles: 1) *Admin*: Manages users, uploads text corpora, assigns tasks, and oversees the corpus development process. 2) *Translator*: Translates the text/corpus into other specified languages. 3) *Reviewer*: Reviews and validates the translations provided by translators.

The system manages the structured data, including metadata related to corpus files such as domain or category, source, and timestamps. The overall workflow of the application has been outlined in Figure 1.

3 Functionalities

3.1 Admin Features

The admin module supports role-based user management under a single super admin, enabling account creation, updates, deletion, password resets,

¹<http://zanata.org/>

²<http://www.teitok.org/>

³<https://pootle.translatehouse.org/>

and blocking for Translator, Reviewer, and Admin level users. All the actions are logged with a timestamp, actor, and remarks. Corpus datasets are uploaded as `.txt` files (one sentence per line), auto-assigned unique IDs per sentence and organized into traceable batches, with full view, update and delete support, bulk/individual operations, versioned history, and timeline-based recovery. Admin workflows allow marking data ready for translation and finalizing reviewed data into the main corpus. The system supports dynamic language addition and a dashboard providing metadata, lexical density, per-user performance reports, workflow oversight, and progress monitoring.

3.2 Translator Features

There are three types of translators: 1) *Crowd*, 2) *Translator*, and 3) *Machine*. *Crowd* translators are registered users who sign up through the platform and contribute translations voluntarily. *Translator* users are professional translators added and managed by the system administrator. *Machine* translators represent automated translation services, such as machine learning models or online translation APIs, which are used when available for a language pair (e.g., Nepali ↔ English). Users can translate each sentence available in their preferred language, with the preference set in the translator’s profile. The system supports tracking work progress, provides reviewer status on their own translation, and maintains a history of edits of translations with version numbers. The edits can be recovered as needed.

3.3 Reviewer Features

Reviewers review the translations provided by the translators, make corrections if needed while maintaining the version of edits, allow the option to skip the review, maintain flags such as “Doubtful”, “Language”, and “Gender”, where the flag can be dynamic and is maintained in the database so that group-based reporting can be done, and track the work progress.

4 Technologies Used

The system is implemented as a client/server web-based application that can be hosted on a web server (local or cloud) and accessible via standard web browsers. We used a modular architecture with a Python-based (Flask) backend, and a PostgreSQL database to store corpus sentences, corpus

metadata, translations, and user information.

5 Application Availability and Licensing

The *PCDT* is developed as part of an effort to build a machine translation system for the Nepali-Tamang language (Ghimire et al., 2026). The source code⁴ and documentation are available for research purposes.⁵

6 Conclusion

The Parallel Corpus Development Toolkit (*PCDT*) provides a robust, scalable solution for building a trilingual (and potentially multilingual) parallel corpus with Nepali as the primary language. By supporting batch uploads, dynamic language additions, and seamless collaboration between *Admins*, *Translators*, and *Reviewers*, the system will play a vital role in developing the parallel corpus required for training high-quality Neural Machine Translation.

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

- 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.
- 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. NepTam: A Nepali-Tamang Parallel Corpus and Baseline Machine Translation Experiments. *arXiv preprint arXiv:2603.14053*.
- Yimam, Seid Muhie, Iryna Gurevych, Richard Eckart de Castilho, and Chris Biemann. 2013. WebAnno: A flexible, web-based and visually supported system for distributed annotations. In Butt, Miriam and Sarmad Hussain, editors, *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 1–6, Sofia, Bulgaria, August. Association for Computational Linguistics.

⁴Sources and documentations: <https://github.com/ilprl/pcdt>

⁵The tool is currently in production and hosted at <https://cat.ilprl.ku.edu.np/>