



Nepalí

Parallel Text Corpus:

Linguistic Features and Structures



NEPALI PARALLEL TEXT CORPUS: LINGUISTIC FEATURES AND STRUCTURES



34

Annotated, quality language data (both-text & speech) and tools in Indian Languages to Individuals, Institutions and Industry for Research & Development - Created in-house, through outsourcing and acquisition.

Authors:

*Umesh Chamling Rai,
Dr. Rejitha K. S.,
Dr. Narayan Choudhary,
Prof. Shailendra Mohan*

**Linguistic Data Consortium for Indian Languages
Central Institute of Indian Language
Mysuru, India-570006**

भारतीय भाषा संस्थान

उच्चतर शिक्षा विभाग, शिक्षा मंत्रालय, भारत सरकार, मानसगंगोत्री, हुन्सूर रोड, मैसूरु-570006, कर्नाटक

CENTRAL INSTITUTE OF INDIAN LANGUAGES

Department of Higher Education, Ministry of Education, Government of India
Manasagangotri, Hunsur Road, Mysuru-570006, Karnataka

www.ciil.gov.in :: e-✉ ciil.publication@gmail.com :: ☎ +91-821-234-5182

Nepali Parallel Text Corpus: Linguistic Features and Structures

Authors: Umesh Chamling Rai, Dr. Rejitha K. S., Dr. Narayan Choudhary, Prof. Shailendra Mohan

Contributors: Sharmila Sharma, Padam Chhetri, Rupesh Rai

e-ISBN: 978-81-69099-73-8

CIIL Publication No.: 1591

First published: AD 2026 March :: Phalguna 1947 Śaka

© Central Institute of Indian Languages, Mysuru 2026

Publisher: Prof. Shailendra Mohan, Director, CIIL

Printing & Publication Unit

Aleendra Brahma, Officer-in-Charge, Printing Press & Publication Unit
B Manju Bhargavi, Unit-in-Charge, Printing Press & Publication Unit
M.N. Chandrashekar, Compositor, Printing Press & Publication Unit

Cover design: Umesh Chamling Rai

Table of Contents

Nepali Parallel Text Corpus.....	1
1.1 LDC-IL Parallel Text Corpus.....	1
1.2 Challenges in Creating A Parallel Corpus.....	2
1.3 Nepali: A Brief Introduction	2
1.4 Nepali parallel Corpus.....	3
1.5 Summary of Nepali parallel Corpus.....	3
1.6 Reference.....	4

NEPALI PARALLEL TEXT CORPUS

Umesh Chamling Rai

1.1 LDC-IL PARALLEL TEXT CORPUS

LDC-IL has developed parallel corpus in Indian mother tongues. India has 270 mother tongues as per 2011 census. Most of these mother tongues have no written text readily available and they are mostly oral languages. Most of them do not have any script of their own or use one or more than one dominant script used in that region. Following the requirements of the NEP-2020, LDC-IL initiated to collect sentences from different mother tongues that could give an idea about the basic structure of these 270 mother tongues.

A parallel corpus is a collection of texts in two or more languages that are carefully aligned sentence wise in pairs and placed either side by side or matched with a common sentence/phrase identification number. It is widely used in linguistics, translation studies, and Natural Language Processing. Researchers use it to compare vocabulary, grammar, and meaning across languages, while computational systems rely on it to train statistical and neural machine translation models. Parallel corpora also support language learning by providing real translation examples and enable cross-language information retrieval, allowing users to access information in multiple languages efficiently.

LDC-IL developed a framework to accommodate all the linguistic features of mother tongues. The sentences were collected from Indian scheduled languages especially from descriptive grammar books. The sentences cover the general grammatical and structural types like adjectival phrases, adverbial phrases, compounding, Interrogative sentences, Imperative sentences, different speech types, statements, structural question, coordination, subordination and so on. Moreover, it includes linguistic features such as use of adposition, anaphora, derivational morphology, inflectional morphology and language specific features like emphasis, equative, possession, reciprocation, reflexive expression and so on. For more details on the source collection for this parallel corpus, please refer to the [LDC-IL Corpus Insights \(2025\)](#).

This parallel corpus can be used in AI development by enabling the creation of Indic multilingual foundation models, develop cross-lingual knowledge graphs and create low-resource language adapters. By providing high-quality linguistic data, it strengthens natural language understanding and generation. As India rapidly advances in the digital age, it is essential to preserve and promote undocumented mother tongues by bringing them onto digital platforms, ensuring they are sustained and able to thrive for future generations.

Low-resource languages have a significant impact on both AI development and digital world. When AI systems are trained mostly on high-resource languages, they become biased toward those linguistic and cultural contexts. Including low-resource languages helps build stronger multilingual foundational models. Exposure to diverse grammar structures, vocabulary, and linguistic patterns improves cross-lingual learning and generalization. Language carries cultural knowledge. Incorporating low-resource languages enriches AI systems with diverse worldviews, idioms, and traditional knowledge systems. If a language is not supported digitally, its speakers

may be excluded from online education, e-governance, healthcare services, and digital commerce. When digital tools support local languages, more people can participate in the economy. Local languages often contain indigenous knowledge about agriculture, medicine, environment, and culture and therefore bringing them into digital ecosystem safeguards the linguistic treasures that language might contain.

Supporting low-resource languages makes AI more inclusive, intelligent, and representative of linguistic diversity. In India, where hundreds of mother tongues co-exist, investing in low-resource language technologies is essential for equitable digital growth.

1.2 CHALLENGES IN CREATING A PARALLEL CORPUS

Creating a Parallel Corpus for mother tongues in India involves several linguistic, technical, and socio-cultural challenges. The major difficulty is the lack of standardized orthography in many regional and tribal mother tongues. Most of the mother tongues use the prominent script used in their region. Some mother tongues may use multiple scripts or even a newly created writing system, which complicates the corpus development. Another challenge is the scarcity of written and digitized resources. Many mother tongues are used for daily communication only.

Dialectal variation is another serious issue, as a single mother tongue may have numerous regional varieties with differences in vocabulary, pronunciation, and grammar. This makes it difficult to ensure consistency and representativeness in the corpus. Accurate translation is a concerned matter. Proficient bilingual speakers may be limited for low-resource languages. Maintaining semantic equivalence across mother tongues with different cultural contexts can be challenging. Sometimes equivalent words are not available due to the socio-cultural differences. These factors affect large-scale parallel corpus development in Indian mother tongues. The creation of reliable and balanced parallel corpora for Indian mother tongues is a complex and resource-intensive task.

1.3 NEPALI: A BRIEF INTRODUCTION

The Nepali language is the official and most widely spoken language in Darjeeling and the state of Sikkim in India. It is belonging to the Indo-Iranian languages branch of the Indo-European language family, and it is native to the Himalayan region of South Asia. This language is spoken in many states of India, including Assam, Manipur, Meghalaya, Arunachal Pradesh, Mizoram, Nagaland, Tripura, Uttarakhand, Jharkhand, Delhi and others. At present, Nepali-speaking communities are also found across the world as part of a global diaspora.

Nepali language has official status in the Gorkhaland Territorial Administration, a semi-autonomous administrative region of West Bengal. More than 80 million people speak this language worldwide. According to the 2011 census report, 29,25,796 Nepali speakers speak this language in India. Nepali is written from left to right using the Devanagari script. The language has a long history, and from ancient times to the modern era it has developed a rich literary tradition and cultural heritage.

1.4 NEPALI PARALLEL CORPUS

Nepali generally follows the Subject–Object–Verb (SOV) word order. In a typical sentence, the subject comes first, the object follows, and the verb appears at the end (काम उ गर्छ/ kamΛ u gartsʰΛ). Nouns in Nepali may show gender (masculine and feminine) and number (singular and plural). Plurality is often indicated by the suffix –हरू (–haru), although plural marking is sometimes optional depending on context. Nepali uses postpositions rather than prepositions. These elements follow nouns and indicate grammatical relationships such as possession, location, direction, and instrument.

Verbs in Nepali are conjugated according to tense, aspect, mood, person, and level of politeness. The language distinguishes between different levels of respect, which influence verb forms and pronouns used in communication. Nepali has a rich pronoun system that reflects social hierarchy and politeness. For example, different forms of 'you' (तँ/taṅ, तिमी/timi, तपाईँ/taṭpaĩ, आफू/apʰu, हजुर/hɪɒdzurΛ, मौसूफ/mausupʰΛ) are used depending on the level of respect shown to the listener. Adjectives in Nepali may agree with nouns in gender and number in some contexts, though many adjectives remain unchanged regardless of the noun.

Nepali uses case markers attached to nouns or pronouns to indicate grammatical roles. Some common case markers are, ले (le) for the agent in past transitive sentences –लाई (lāĩ) for the dative or indirect object–को (ko) for possession –मा (mā) for location. Nepali frequently uses particles such as नै (nai), त (ta), and पो (po) to add emphasis, focus, or nuance to a sentence. Nepali vocabulary includes many words borrowed from Sanskrit, Hindi, and neighboring Tibeto-Burman (नाना/naṇa, ढोका/dʰoka, फाक्सा/pʰaksa, थीं/θiŋ, म्हम्ह/mʰɪɒmʰɪΛ, नाकिमा/ṇakima) language family, reflecting centuries of cultural and linguistic interaction. The above mentioned features contribute to Nepali's uniqueness among the languages of India and highlight its importance in linguistic studies.

Nepali is a valuable resource for developing parallel text corpora with other Indian mother tongues. This parallel text corpus can serve as a valuable research resource for studying syntactic divergence, morphological richness, code-mixing phenomena, and script variation, which are particularly relevant in multilingual contexts such as India.

1.5 SUMMARY OF NEPALI PARALLEL CORPUS

The Nepali parallel text corpus connected with English and 146 mother tongues of India. Those mother tongues are Anal, Angami, Are, Assamese, Awadhi, Bagheli/Baghel Khandi, Bagri Rajasthani, Bagri, Balti, Bengali, Bhadrawahi, Bharmauri/Gaddi, Bhojpuri, Bilaspuri Kahluri, Bodo, Brajbhasha, Bundeli/Bundel khandi, Chakru/Chokri, Chambeali/Chamrali, Chang, Chhattisgarhi, Chirri, Chungli, Churahi, Coorgi/Kodagu, Deori, Dhundhari, Dimasa, Dogri, Gangte, Garhwali, Garo, Gojri/Gujjari/Gujar, Gujarati, Gujar, Halabi, Handuri, Hara/Harauti, Haryanvi, Hindi Multani, Hindi, Irula/Irular Mozhi, Kabui, Kachchhi, Kangri, Kannada,

Kashmiri, Khandeshi, Khari Boli, Khasi, Khezha, Khiemnungan, Khortha/Khotta, Kisan, Kodava, Kolami, Koli, Kom, Konda, Konkani, Koya, Kudubi/Kudumbi, Kuki, Kurmal Thar, Lepcha, Liangmei, Lotha, Lyngngam, Magadhi/Magahi, Maithili, Malayalam, Malvi, Manipuri, Mao, Mara, Maram, Marathi, Maring, Mech/Mechhia, Mewari, Mewati, Miri/Mising, Mishmi, Mongsen, Monpa, Mundari, Muwasi, Nawait, Nimadi, Nyishi, Nocte, Odia, Pahari, Paite, Palmuha, Pania, Pawari/Powari, Phom, Pnar/Synteng, Pochury, Poula, Punjabi, Purkhi, Rai, Rajasthani, Reang, Rengma, Rongmei, Sadan/Sadri, Sambalpuri, Sangtam, Sanskrit, Santali, Saurashtra/Saurashtri, Sema, Shina, Sindhi, Sirmauri, Sugali, Surjapuri, Talgalo, Tamil, Tangkhul, Telugu, Thado, Tikhir, Tripuri, Tulu, Urdu, Vaiphei, Wagdi, Wancho, Yerava, Yerukala/Yerukula, Yimchungre, Zeliang, Zemi, Zou.

There are 5,332 sentences/phrases in Nepali which are systematically structured based on 159 grammatical categories. Nepali has 27397 words count and 155604 character count. The corpus is aligned with 146 mother tongues and English, comprising 4,404,845 words (over 4.4 million tokens) and 23,374,289 characters (approximately 23.3 million).

1.6 REFERENCE

- [1] Rejitha K. S. and Narayan Kumar Choudhary, (ed.). (2025), LDC-IL Corpus Insights. CIIL, Mysore. 978-93- 48633-33-0.
- [2] Ghanashyam Nepal & Kavita Lama, 2018, Uchcha madhyamik Nepali vyakran ra rachana, Ekta Book House, Silgadi, 978-81-8161-398-1.
- [3] Hemangaraj Adhikari, 2015, Samasamayik Nepali Vyakran, Kathmandu, Nepal.
- [4] A. Trunbull, 1904, Nepali grammar and Eng-Nep & Nep-Eng Vocabulary, SMO Press, Darjeeling, India.