Tibetan

Parallel Text Corpus:

Linguistic Features and Structures

TIBETAN PARALLEL TEXT CORPUS: LINGUISTIC FEATURES AND STRUCTURES

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

Tibetan Parallel Text Corpus: Linguistic Features and Structures

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

Contributors: Dr. Lhamo Tso, Lobsang Gendun

e-ISBN: 978-81-69175-68-5
CIIL Publication No.: 1721

First published: AD 2026 March :: *Phalguna 1947 Śaka*

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

TIBETAN 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 TIBETAN: A BRIEF INTRODUCTION

The Tibetan language belongs to the Sino-Tibetan language family, specifically to the Tibeto-Burman language family branch. It is primarily spoken in Tibet, as well as in Ladakh, Arunachal Pradesh, Himachal Pradesh, and various parts of India. It is also spoken in Nepal and Bhutan. Tibetan is an important cultural and religious language, especially in relation to Tibetan Buddhism, where it is widely used for scriptures, philosophy, and monastic education.

Tibetan has approximately 6 to 7 million speakers worldwide. According to the 2011 census report, there are 83,779 Tibetan speakers in India. The Tibetan language is widely used in religious texts, literature, and communication among Tibetan communities. It is written using the Tibetan script, which was developed in the 7th century. Tibetan literature includes religious scriptures, historical chronicles, poetry, and philosophical writings that reflect the spiritual and intellectual traditions of the region. In addition, Tibetan has a rich tradition of classical literature.

1.4 TIBETAN PARALLEL CORPUS

The Tibetan language has several regional varieties, including Central Tibetan, Amdo Tibetan, and Khams Tibetan. These varieties differ in pronunciation, vocabulary, and grammar, and some may not be mutually intelligible. Tibetan typically follows the Subject–Object–Verb (SOV) word order. The language uses postpositions rather than prepositions, and grammatical relationships are often expressed through particles and suffixes. Tibetan also shows ergative features in certain grammatical constructions. Verbs generally indicate tense, aspect, and mood, while nouns may take case markers to show their grammatical role in a sentence.

Tibetan shows agglutinative characteristics, meaning grammatical information is expressed through the addition of suffixes or particles to a base word. It is also known for its system of honorific forms, which are used to show respect when speaking to or about people of higher status, religious figures, or elders. Nouns in Tibetan can take case markers that indicate their grammatical role in a sentence, such as agent, object, location, or possession. These markers function similarly to postpositions.

Tibetan verbs do not change extensively for person or number. Instead, tense and aspect are expressed through auxiliary verbs and particles. Verbs may also have different stems to indicate past, present, or future meaning. Their grammar relies heavily on particles rather than inflection. These particles indicate grammatical relationships such as case, tense, emphasis, and mood. They are usually placed after nouns or verbs. Nominalization is common in Tibetan.

Tibetan has a unique writing system. Compared to other languages, the word count in Tibetan is often five to six times lower. In most languages, a sentence typically consists of a combination of two, three, four, or more words. However, due to the writing system of Tibetan, many words within a single sentence can combine to form a single word. Therefore, the number of words in Tibetan sentences is often much lower than in other languages. This phenomenon also occurs in Ladakhi. With this the above mentioned features contribute to Tibetan's uniqueness among the languages of India and highlight its importance in linguistic studies.

Tibetan 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 TIBETAN PARALLEL CORPUS

The Tibetan 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, Chirr, Chungli, Churahi, Coorgi/Kodagu, Deori, Dhundhari, Dimasa, Dogri, Gangte, Garhwali, Garo, Gojri/Gujjari/Gujar, Gujarati, Gujari, 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, Kurmali 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 Tibetan which are systematically structured based on 159 grammatical categories. Tibetan has 5497 words count and 176021 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] Sarat Chandra Das, 1915, Introduction to the Grammar of the Tibetan Language with the texts of Situ Sumtag, Darjeeling, India.