Sanskrit Parallel Text Corpus:

Linguistic Features and Structures

SANSKRIT 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:

Chetan Baji

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

Sanskrit Parallel Text Corpus: Linguistic Features and Structures

Authors: Chetan Baji, Dr. Rejitha K. S., Dr. Narayan Choudhary, Prof. Shailendra Mohan

Contributors: Dr. Vinayak Bhat, Varun R, Dr.Vishwanath M V

e-ISBN: 978-81-69175-47-0
CIIL Publication No.: 1735

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: Chetan Baji

Table of Contents

SANSKRIT PARALLEL TEXT CORPUS

Chetan Baji

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 SANSKRIT: A BRIEF INTRODUCTION

Sanskrit is an ancient Indo-Aryan classical language that has profoundly shaped the intellectual, spiritual, and cultural traditions of South Asia for over three millennia. With roots dating back more than 3,500 years, its earliest recorded form appears in the Rigveda, the oldest extant Indo-Aryan text. Sanskrit is traditionally regarded as *Devabhāṣā*—the "language of the gods"—because of its sacred association with religious literature and ritual practice.

Sanskrit evolved from earlier Indo-Aryan dialects during the Vedic period and later developed into Classical Sanskrit. The formal structure of Classical Sanskrit was codified by the great grammarian ***Panini*** in his monumental work, the ***Ashtadhyayi***, composed around the 6th–5th century BCE. This treatise, consisting of nearly 4,000 concise rules (sutras), provides a remarkably scientific and generative description of phonetics, morphology, and syntax. ***Panini***'s grammatical system is considered one of the most sophisticated linguistic frameworks in the world and continues to attract attention in modern linguistics and computational studies.

Structurally, Sanskrit is known for its phonetic precision and logical organization. Traditionally written in the Devanagari script (though historically preserved in various regional scripts), it has a highly systematic alphabet comprising 52 letters—16 vowels and 36 consonants—arranged according to phonetic principles. Its rich system of declensions, conjugations, and compound formation allows for concise yet expressive communication.

Sanskrit occupies a central place in world literature. It is the language of the great Indian epics, the ***Mahabharata*** and the ***Ramayana***, as well as the Vedas and the Upanishads. Beyond religious texts, it has produced extensive works in poetry, drama, philosophy, medicine (Ayurveda), astronomy, mathematics, and political thought.

Culturally, Sanskrit is foundational to Indian heritage, deeply influencing yoga, classical dance, music, and philosophical traditions. Sanskrit is recognized as a classical language, Sanskrit continues to serve as a living link between India's ancient past and its contemporary intellectual life.

1.4 SANSKRIT PARALLEL CORPUS

Sanskrit is a highly inflectional and morphologically rich language characterized by a well-defined grammatical system and remarkable structural precision. It exhibits complex morphophonemic processes governed by the rules of ***Sandhi***, an elaborate case system with eight grammatical cases, three numbers (singular, dual, plural), and three genders. Verbal morphology is equally sophisticated, marked by tense–aspect–mood distinctions, voices (parasmaipada and ātmanepada), and a wide range of derivational processes including primary and secondary suffixation (kṛt and taddhita formations). Compounding (samāsa) is highly productive and allows the formation of semantically dense expressions.

Sanskrit is traditionally written in the Devanagari script, though historically it has been represented in various regional scripts across India. Its phonological system is scientifically organized, with a systematic classification of vowels and consonants based on place and manner of articulation. While Sanskrit generally follows a Subject–Object–Verb (SOV) word order, its rich inflectional morphology permits considerable syntactic flexibility without loss of clarity.

These linguistic features contribute to Sanskrit's uniqueness and establish it as one of the most systematically described classical languages in the world, especially through the grammatical tradition of ***Paṇini*** and his foundational work, the ***Ashṭadhyayi***. Sanskrit holds immense importance in linguistic theory, comparative philology, and the study of Indo-European languages.

Sanskrit also serves as a valuable resource for developing parallel text corpora with other Indian mother tongues. A Sanskrit-based parallel corpus can significantly contribute to research in areas such as syntactic alignment, morphological complexity, derivational productivity, semantic density in compounds, and cross-script representation. Such corpora are particularly relevant in

multilingual contexts like India, where Sanskrit has historically influenced numerous regional languages at lexical, phonological, and syntactic levels.

This Sanskrit parallel corpus constitutes a systematically translated dataset derived from an original English corpus, ensuring structured cross-linguistic alignment and semantic correspondence.

1.5 SUMMARY OF SANSKRIT PARALLEL CORPUS

The Sanskrit parallel text corpus is aligned with English and 146 mother tongues of India. And These mother tongues are Anal, Angami, Apatani, 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, Karbi/Mikir, Kashmiri, Khandeshi, Khari Boli, Khasi, Khezha, Khiemnungan, Khortha/Khotta, Kisan, Kodava, Kokbarak, Kolami, Koli, Kom, Konda, Konkani, Konyak, Koya, Kudubi/Kudumbi, Kuki, Kurmali Thar, Ladakhi, Lepcha, Liangmei, Limbu, Lotha, Lyngngam, Magadhi/Magahi, Maithili, Malayalam, Malvi, Manipuri, Mao, Mara, Maram, Marathi, Maring, Mech/Mechhia, Mewari, Mewati, Miri/Mising, Mishmi, Mizo, Mongsen, Monpa, Mundari, Muwasi, Nawait, Nepali, Nimadi, Nissi, Nocte, Odia, Pahari, Paite, Palmuha, Pania, Pawari/Powari, Phom, Pnar/Synteng, Pochury, Poula, Punjabi, Purkhi, Rai, Rajasthani, Reang, Rengma, Rongmei, Sadan/Sadri, Sambalpuri, Sangtam, Santali, Saurashtra/Saurashtri, Sema, Shina, Sindhi, Sirmauri, Sugali, Surjapuri, Talgalo, Tamil, Tangkhul, Telugu, Thado, Tibetan, Tikhir, Tripuri, Tulu, Urdu, Vaiphei, Wagdi, Wancho, Yerava, Yerukala/Yerukula, Yimchungre, Zeliang, Zemi, Zou.

There are 5,332 sentences/phrases in Sanskrit which are systematically structured based on 159 grammatical categories. Sanskrit has 24,136 word count and 152,626 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] Deshpande, M. M. (1993). *Sanskrit and Prakrit: Sociolinguistic issues*. Delhi: Motilal Banarsidass.
[3] Sharma, R. S. (2002). *India's ancient past*. New Delhi: Oxford University Press.
[4] Ruppel, A. M. (2017). *The Cambridge introduction to Sanskrit*. Cambridge: Cambridge University Press.