

Assamese Parallel Text Corpus:

LINGUISTIC FEATURES AND STRUCTURES

ভাৰতীয়

অসমীয়া



ASSAMESE 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:

Syeda Mustafiza Tamim

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

Assamese Parallel Text Corpus: Linguistic Features and Structures

Authors: Syeda Mustafiza Tamim, Dr. Rejitha K. S., Dr. Narayan Choudhary, Prof. Shailendra Mohan

Contributors: Angelina B. Dihingia, Priyanshe Adhyapak, Bikash Chetia, Puja Das, Udipta Saikia

e-ISBN: 978-81-69099-76-9

CIIL Publication No.: 1682

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: Syeda Mustafiza Tamim

Table of Contents

Assamese Parallel Text Corpus	1
1.1 LDC-IL Parallel Text Corpus.....	1
1.2 Challenges in Creating A Parallel Corpus.....	2
1.3 Assamese: A Brief Introduction	3
1.4 Assamese parallel Corpus	4
1.5 Summary of Assamese parallel Corpus	4
1.6 Reference.....	5

ASSAMESE PARALLEL TEXT CORPUS

Syeda Mustafiza Tamim

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 systems, 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 ASSAMESE: A BRIEF INTRODUCTION

Assamese is the easternmost Indo-Aryan language, spoken by over **15 million people**. It is the official language of Assam and serves as a vital medium of communication, education, administration, literature, and cultural expression. Linguistically, Assamese belongs to the Eastern Indo-Aryan branch and shares historical roots with Bengali and Odia, though it has developed distinct phonological, morphological, and syntactic features. The language uses the Assamese script, a variant of the Eastern Nagari script, and has a rich literary heritage dating back to the medieval period.

Assamese is not only a language but also a strong marker of regional identity and cultural continuity. It is a lingua franca among diverse communities of Assam. The language has a vibrant tradition of poetry, fiction, drama, journalism, and modern digital media. The 14th-century poet Madhav Kandali is known for composing one of the earliest vernacular translations of the Ramayana into Assamese. Later, the 15th–16th century saint-scholar Srimanta Sankardev played a transformative role in shaping Assamese literature, drama, music, and devotional culture through the Neo-Vaishnavite movement. Later prominent literary figures such as Lakshminath Bezbaruah and Homen Borgohain contributed significantly to the growth of modern Assamese prose and national consciousness. In contemporary times, the language is actively used in academic research, computational linguistics, and language technology initiatives, supporting corpus development, speech processing, and preservation of linguistic resources. Thus, Assamese stands as a dynamic and evolving language that reflects the historical depth, cultural richness, and intellectual vitality of the people of Assam.

1.4 ASSAMESE PARALLEL CORPUS

Assamese is an Indo-Aryan language moderately inflectional morphology with agglutinative tendencies, as grammatical information is typically expressed through suffixes attached sequentially to the root word and systematic case marking. It uses classifiers, definiteness markers, and aspectual distinctions that add structural depth to sentence formation. Although the basic word order is Subject–Object–Verb, discourse factors allow flexibility. Written in the Assamese script derived from the Brahmi tradition, the language reflects centuries of contact among diverse communities in Assam, giving it distinctive phonological and lexical features within India.

An Assamese parallel corpus with other Indian languages is highly valuable for comparative linguistic and computational research. It supports the study of grammatical variation, translation patterns, and morphological complexity across languages. Such a resource is essential for developing machine translation, cross-lingual tools, and NLP applications, while also strengthening digital documentation and preservation of Assamese in multilingual contexts.

1.5 SUMMARY OF ASSAMESE PARALLEL CORPUS

The Assamese parallel text corpus connected with English and 146 mother tongues of India. Those mother tongues are Anal, Angami, Apatani, Are, 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, Sanskrit, 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 Assamese which are systematically structured based on 159 grammatical categories. Assamese has 26,055 word count and 147,249 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] Encyclopaedia Britannica. (n.d.). Assamese language.
- [3] Ethnologue / Census of India data cited via Wikipedia (2011).
- [4] Languages of India. (n.d.). Assamese language section.
- [5] ITIHAAS. (n.d.). Assamese language status and speakers
- [6]https://language.census.gov.in/eLanguageDivision_VirtualPath/MTSI/ASSAM/Assamese.pdf