# Angami Parallel Text Corpus: Linguistic Features and Structures

# ANGAMI 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:**

***Yumnam Premila Chanu***
***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

Angami Parallel Text Corpus: Linguistic Features and Structures

*Authors:* Yumnam Premila Chanu, Dr. Rejitha K. S., Dr. Narayan Choudhary, Prof. Shailendra Mohan

*Contributors:* Rüübino Peseyie, Thepfulenuo Mere, Vetsino, Vibeituonuo Mere

*e-ISBN: 978-81-69099-40-0*
*CIIL Publication No.: 1685*

*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:* Sonali Sutradhar

# Table of Contents

# ANGAMI PARALLEL TEXT CORPUS

*Yumnam Premila Chanu*

## 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 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 ANGAMI: A BRIEF INTRODUCTION

Angami (locally known as Tenyidie) is a prominent Sino-Tibetan language spoken primarily in the Kohima and Dimapur districts of Nagaland, Northeastern India, by the Angami Naga community. Belonging to the Angami-Pochuri branch of the Tibeto-Burman family, it serves as a crucial linguistic bridge and a secondary language for several neighboring Naga groups. According to the 2011 Census of India, the Angami-speaking population is approximately 152,000, underscoring its status as one of the most widely spoken and influential indigenous languages in the region.

The Angami language is written using the Latin (Roman) script, a transition that began in the late 19th century under the influence of American Baptist missionaries. This writing system was later standardized into Tenyidie, which now serves as the official medium for education, administration, and literature. A defining characteristic of Angami is its complex tonal system, which is among the most sophisticated in the Sino-Tibetan family.

Angami literature is a rich tradition that began with oral narratives, including ancient myths, legendary folktales, and folksongs that have passed down the history and values of the Angami peoples for generations. The transition to written Angami literature started in the late 19th and early 20th centuries with the introduction of the Latin alphabet. Christian missionaries, local teachers, and scholars played a crucial role in this process by documenting the language and creating early educational and religious texts. Today, Angami writers produce a wide range of original works, including: Poetry and Prose, Short Stories and Essays, and Historical Writings. Angami literature plays a vital role in preserving the Angami identity. It serves as a bridge for younger generations, helping them stay connected to their history, traditional knowledge, and the unique cultural heritage of their ancestors

## 1.4 ANGAMI PARALLEL CORPUS

Angami is characterized by its highly agglutinative nature, featuring a sophisticated system of verbal morphology. The language utilizes elaborate sequences of suffixes to modify roots, creating a layer of structural complexity that is central to its grammar. Adhering to the standard framework of the Sino-Tibetan family, Angami follows a consistent Subject-Object-Verb (SOV) word order. A defining phonological trait of the language is its complex tonal system, which typically features several distinct pitch levels.

Angami serves as a vital resource for creating parallel text corpora alongside other Indian regional and mother tongue languages. It offers unique datasets for researchers investigating syntactic divergence, morphological richness, and the dynamics of code-mixing- phenomena that are especially prevalent in the diverse, multilingual landscape of Northeast India. Furthermore, since Angami utilizes an adapted Latin script with specific orthographic conventions to represent its varied tones, it provides an excellent case study for script variation and the evolution of writing systems for indigenous tribal languages.

## 1.5 SUMMARY OF ANGAMI PARALLEL CORPUS

The Angami parallel text corpus connected with English and 146 mother tongues of India. Those mother tongues are Anal, 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, 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 Angami which are systematically structured based on 159 grammatical categories. Angami has 28527 word count and 150397 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.