Apatani Parallel Text Corpus:

Linguistic Features and Structures

APATANI 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

Apatani Parallel Text Corpus: Linguistic Features and Structures

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

Contributors: Bamin Yalung, Bamin Lure

e-ISBN: 978-81-69099-46-2
CIIL Publication No.: 1684

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

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

APATANI 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 APATANI: A BRIEF INTRODUCTION

The Apatani language, also known as Tani Agun, belongs to the Tani branch of the Sino-Tibetan (Tibeto-Burman) language family. It is primarily spoken by the Apatani people of India. Unlike many neighboring languages, Apatani is concentrated within a specific geographic plateau, which has allowed it to develop distinctive characteristics separate from other Tani languages such as Nyishi and Adi.

According to the 2011 census report the number of Apatani speakers is 44815 in India. Although the community is relatively small, the language remains vibrant and plays a central role in the cultural identity of the Apatani people in the valley. Interestingly, because the Apatani population is concentrated within a single valley, the language does not have major regional dialects that hinder mutual intelligibility. Speakers from different villages can understand one another easily. However, there are subtle clan-based or village-based variations in pronunciation and certain vocabulary items. These lexical differences occur among villages such as Hong, Hari, Bulla, Hija, Dutta, and Bamin-Michi. The main difference lies in the fact that each clan may

have its own way of pronouncing or expressing certain words. These variations are generally phonetic or stylistic rather than structural.

1.4 APATANI PARALLEL CORPUS

Apatani follows a Subject–Object–Verb (SOV) sentence structure. It is an agglutinative language, meaning that grammatical information is expressed through the addition of suffixes to root words. The language also uses a classifier system, where numbers must be paired with classifiers based on the shape or category of the object (for example, flat objects, long objects, or humans).

Another important feature of Apatani is the role of phonetic pitch, which can influence meaning and therefore requires precision in both speech and its standardized Roman script orthography. The language often uses suffixes that indicate the source of the speaker's knowledge, such as whether the information was personally witnessed or heard from someone else.

Unlike English pronouns, which are gendered (he/she), Apatani pronouns are generally gender-neutral but socially stratified. This means that translators often need to infer the social relationship between speakers in order to choose the appropriate level of respect or formality.

Furthermore, Apatani does not always have a one-to-one lexical correspondence with source languages such as English, particularly for modern technical, legal, or administrative terms. While English contains a large number of technical and abstract nouns, Apatani vocabulary is deeply rooted in the natural environment and social relationships of the community.

To interpret simple Hindi or English terms such as 'brother' or 'sister' requires greater specificity in Apatani. The language distinguishes between elder and younger siblings and often reflects social hierarchy. Therefore, interpreters must infer the context of the source sentence in order to choose the correct Tani term. Another important linguistic challenge involved managing sentence length. Because Apatani is an agglutinative language, a long English sentence containing multiple prepositions can sometimes be compressed into a much shorter Apatani sentence, where the verb carries most of the grammatical information. Conversely, abstract English nouns often require longer, descriptive Apatani clauses in order to convey the intended meaning accurately.

Although Apatani is not as strictly tonal as some other languages, the flow and pitch of a sentence can influence the intended meaning. Therefore, ensuring that polite forms were used correctly during the translation of official instructions was essential in order to maintain the dignity and appropriateness of the language.

In addition, formal English writing frequently uses the passive voice, which often sounds unnatural in Apatani. As a result, many sentences had to be converted into the active voice to preserve naturalness and fluency in the target language. The above mentioned features contribute to Apatani's uniqueness among the languages of India and highlight its importance in the study of linguistics.

Apatani 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 APATANI PARALLEL CORPUS

The Apatani 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, 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, 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 Apatani which are systematically structured based on 159 grammatical categories. Apatani has 33911 word count and 167364 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] E Annamalai (ed.). (1985). Apatani Grammar. CIIL, Mysore.

[3] P. T. Abraham, Apatani Grammar. CIIL, Mysore.