Dimasa Parallel Text Corpus:

Linguistic Features and Structures

DIMASA 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

Dimasa Parallel Text Corpus: Linguistic Features and Structures

Authors: Yumnam Premila Chanu, Dr. Rejitha K. S., Dr. Narayan Choudhary, Prof. Shailendra Mohan
Contributors: Dr. Monali Longmailai, Dr. Krithika Barman

e-ISBN: 978-81-69099-87-5
CIIL Publication No.: 1658

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

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

DIMASA 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 DIMASA: A BRIEF INTRODUCTION

Dimasa is a language spoken by an ethnically minority community in Assam, India. And it is known by the autonym Graodima is a language of the Bodo-Garo subfamily of the Tibeto-Burman language group under the Sino-Tibetan language family, which is the largest language family in Asia. Most of the Dimasa people live in Assam, especially in the district of Dima Hasao, which was earlier known as North Cachar Hills. However, many members of the community are also found in nearby states like Nagaland, Meghalaya, Manipur, and Mizoram. Even though there are over 111,000 Dimasa speakers in India according to the 2011 Census, the language still has to deal with the challenges of being a minority group surrounded by dominant regional languages. The Hasao, Hawar, Dembra, and Dijuwa are among the several dialectal varieties of the Dimasa language. The Hasao dialect is commonly acknowledged as the as the standard variant used for formal communication and literature among them. Dimasa lacks a traditional indigenous alphabet, in contrast to many of India's major languages; yet, the community has effectively adopted the Roman script for modern writing, education, and literary production.

Dimasa literature represents a vital repository of the history, cosmology, and social values of the Dimasa people. Historically, the language- a member of the Bodo-Garo branch of the Tibeto-Burman family relied on a sophisticated oral tradition to transmit knowledge across generations. For centuries, Dimasa culture was preserved through the Orality of the Kachari (the historical identity of the Dimasa). This folk literature is categorized into several functional genres i.e., Folk Tales and Myths, Folk Songs, Ritual Chants and Incantations, and Proverbs and Riddles. The shift toward written Dimasa was accelerated by the adoption of the Latin (Roman) alphabet, though the Bengali script was also historically used in certain regions. This transition allowed for the systematic documentation of the language and the birth of modern literary genres, it includes Standardization of Language, Creative Prose and Poetry, Dramatic Works etc.

1.4 DIMASA PARALLEL CORPUS

Dimasa is defined by its agglutinative structure, a characteristic shared by many languages within the Bodo-Kachari group. The language features a sophisticated system of verbal morphology and a diverse set of postpositional markers that dictate grammatical relationships. A notable phonological trait of Dimasa is its tonal system, which works alongside complex morphophonemic changes to distinguish meaning between words. Additionally, Dimasa employs a robust collection of numeral classifiers, which categorize nouns based on their physical attributes such as shape, size, and function. While the language strictly follows a Subject-Object-Verb (SOV) pattern, its clear case-marking system allows for some flexibility in sentence structure. These unique features highlight Dimasa’s importance within the broader study of Sino-Tibetan linguistics and its distinct place in Northeast India.

As a tonal language with a sophisticated system of distinct tones, Dimasa offers unique data for researchers studying how pitch changes influence word meanings. This makes it a valuable research resource for examining syntactic divergence, morphological richness, code-mixing phenomena, and script variation (particularly through its standardized Tenyidie form). These studies are especially important in the diverse, multilingual environment of Northeast India, where Dimasa acts as a key representative of the Tibeto-Burman language family.

1.5 SUMMARY OF DIMASA PARALLEL CORPUS

The Dimasa parallel text corpus connected with English and 146 mother tongues of India. Those 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, 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 Dimasa which are systematically structured based on 159 grammatical categories. Dimasa has 23681 word count and 152797 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.