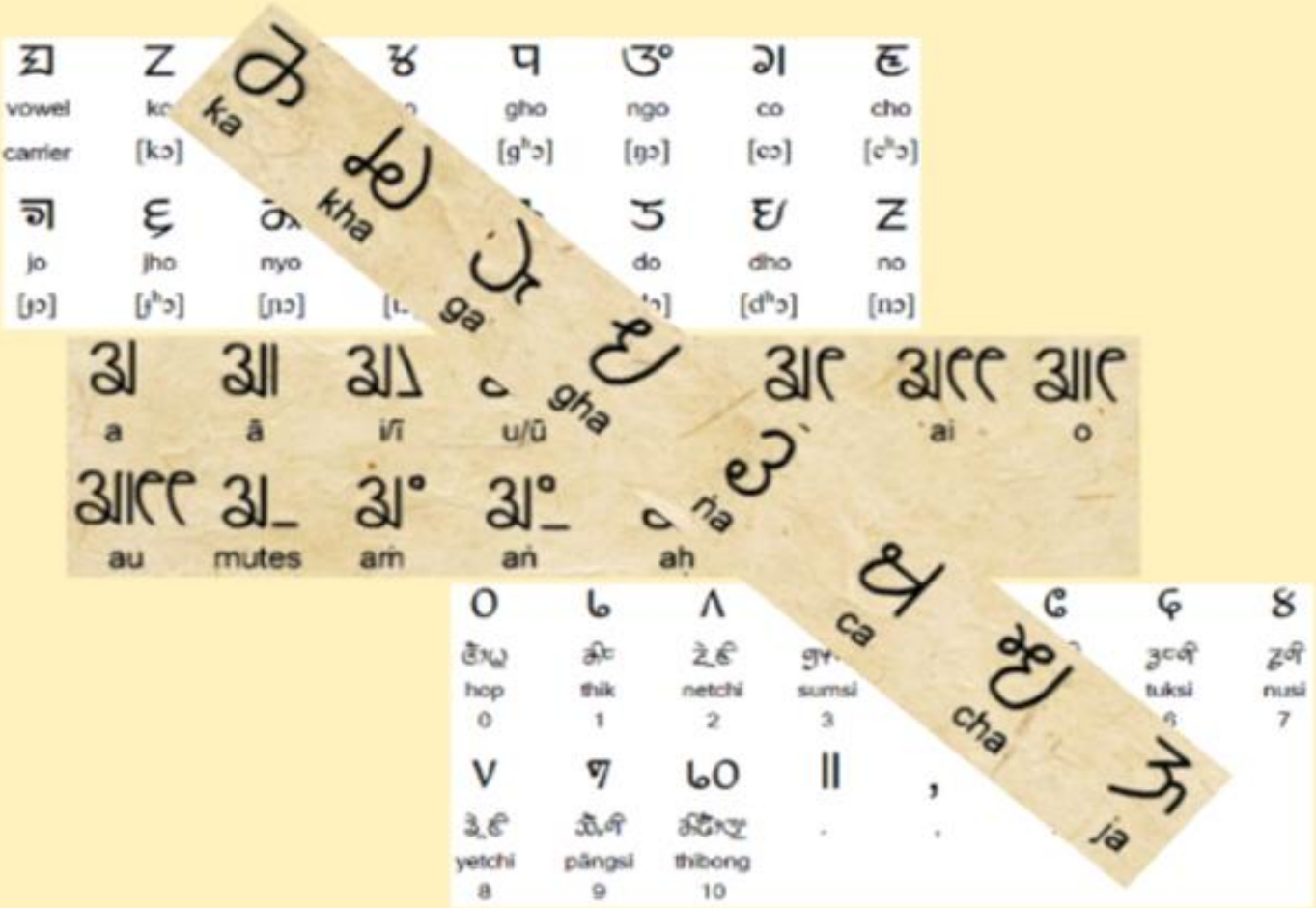
Rai
Parallel Text Corpus:
Linguistic Features and Structures



RAI 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

Rai Parallel Text Corpus: Linguistic Features and Structures

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

Contributors: Taraman Rai, Dil Kumar Rai, Bir Bahadur Rai, Pratima Rai

e-ISBN: 978-81-69175-71-5
CIIL Publication No.: 1743

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

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

RAI 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 RAI: A BRIEF INTRODUCTION

Rai belongs to the Tibeto-Burman language family, under the Kiranti languages. They are also known by Kirat Rai. Rai are one of the indigenous communities settled in several villages of Sikkim, Darjeeling, Dooars region, and other parts of Northeast India. The Rai population in India is estimated to be approximately 600,000, but only a small proportion of them still speak their mother tongue. Many speakers are gradually shifting toward neighboring languages due to opportunities related to jobs, careers, and social mobility. Rai native speakers often refer to their language as Kirawajen or Sawajen. The Government of Sikkim recognized Rai as one of the official languages of the state on 21 March 1995. A large number of Rai language speakers are also live in Nepal.

Rai people have rich traditions, oral literature, and unique religious practices. They follow Kirat Mundhum, an indigenous belief system based on oral scriptures known as the Mundhum, their cultural life includes traditional music, dance, rituals, and festivals. One of their most important festivals is Sakela, which is celebrated with community dances and rituals to honor nature and ancestors. This cultural identity, languages, and traditions remain an important part of the

multicultural heritage of the Himalayan region. Today, efforts are being made by community organizations and scholars to preserve Rai languages and promote their cultural heritage.

1.4 RAI PARALLEL CORPUS

Rai is spoken especially in the Himalayan regions of the southern part of the Asian continent. It is an umbrella term, and within it many other languages are spoken. However, these languages are closely related to one another. Some of them, Bantawa, Chamling, Kulung, Thulung, Sampang, Dumi, Khaling, Wambule, Puma, Koyu, Yamphu, Chulung, Mewahang, Tilung, Sunuwar, Yakha and others. Since Bantawa is understood by most people among them, its use has become more widespread than the others.

Rai has thirty consonant phonemes- /p/, /b/, /t/, /d/, /ʈ/, /ɖ/, /c/, /ɟ/, /k/, /g/, /pʰ/, /bʰ/, /tʰ/, /dʰ/, /ʈʰ/, /ɖʰ/, /cʰ/, /ɟʰ/, /kʰ/, /gʰ/, /m/, /n/, /ɲ/, /ŋ/, /r/, /s/, /j/, /w/, /l/. It consists of six vowels, they are: /i/, /e/, /ə/, /a/, /o/ and /u/. It is noticed that Rai language has aspirated voiceless stop. The consonant phonemes /gʰ/, /ɲ/ and /ŋ/ has been found in this language which is rare in other Tibeto-Burman language. Rai people use /g/, /gʰ/ and /ɲ/ only in few words like /gəkwa/ ‘crow’, /mag/ ‘see’, /gʰolʈoŋ/ ‘cart’, /tuŋlagʰum/ ‘genius’ and /ɲaŋcəŋ/ ‘coins garland, /ɲa/ ‘used as dative case’. Rai language has more suffixes than prefixes; for example, /ba/, /pa/, /wa/, /ma/, /kop/, /mi/, /ula/, /ci/, /cu/, /nin/ etc.

Rai language is well known in linguistics for its highly complex verb morphology. Verbs can contain several prefixes and suffixes that express tense, aspect, mood, person, number, and sometimes both the subject and the object. Because of this, a single verb form may carry a large amount of grammatical information.

It follows the Subject–Object–Verb (SOV) word order. In this structure, the subject comes first, followed by the object, and the verb appears at the end of the sentence. Rai is agglutinative language, meaning words are formed by adding several suffixes to a root word. These suffixes indicate grammatical information such as tense, aspect, person, number, and case. One of the most distinctive features of Rai language is its complex verb agreement system. Verbs often agree with both the subject and the object in a sentence. This system can encode person, number, and sometimes inclusivity.

Nouns are marked with case suffixes that indicate their grammatical role in a sentence, such as subject, object, location, or possession. Instead of prepositions, Rai language usually uses postpositions that appear after nouns. Various particles are also used to express emphasis, focus, or grammatical relations. Rai language use classifiers when counting nouns, especially for objects, animals, or people, it categorizes nouns based on shape, function, or animacy. (In few cases). The above mentioned features contribute to Rai’s uniqueness among the languages of India and highlight its importance in the study of linguistics.

Rai 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 RAI PARALLEL CORPUS

The Rai 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, 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 Rai which are systematically structured based on 159 grammatical categories. Rai has 25213 word count and 173937 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] Doornenbal, M. A. (2009, November 3). A Grammar of Bantawa : grammar, paradigm tables, glossary and texts of a Rai language of Eastern Nepal. LOT dissertation series. LOT, Netherlands Graduate School of Linguistics, Utrecht.
[3] Karen Ebert, 1997, Camling (Chamling), LINCOM EUROPA, Munchen - Newcastle.