

HINDI PARALLEL TEXT CORPUS: LINGUISTIC FEATURES AND STRUCTURES

चिंता मत करो।

वह मेरा दोस्त है।

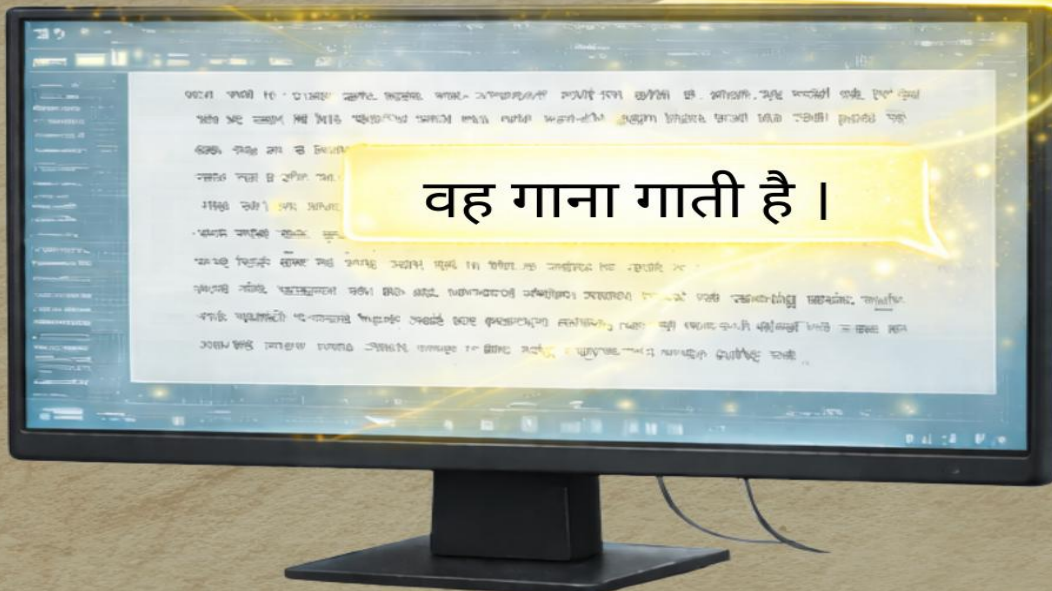
यह बहुत आसान है।

मुझे भूख लगी है।

मैं कल आऊँगा।

आज मेरी मीटिंग है।

वह गाना गाती है।



HINDI 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:

Dr. Satyaendra Kumar Awasthi

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 वभाग, 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

Hindi Parallel Text Corpus: Linguistic Features and Structures

Authors: Dr. Satyaendra Kumar Awasthi, Dr. Narayan Choudhary, Prof. Shailendra Mohan

Contributors: Neha Dixit

e-ISBN: 978-81-69099-96-7

CIIL Publication No.: 1645

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: Teena Verma

Table of Contents

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

HINDI PARALLEL TEXT CORPUS

Dr. Satyaendra Kumar Awasthi

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 HINDI: A BRIEF INTRODUCTION

Hindi is an Indo-Aryan language that evolved from Sanskrit and is widely spoken across central and northern India, particularly in the states of Bihar, Chhattisgarh, Delhi, Haryana, Himachal Pradesh, Jharkhand, Madhya Pradesh, Rajasthan, Uttarakhand, and Uttar Pradesh. It functions as the official language of the Union of India and also serves as a common link language throughout the country. According to the Census of India 2011, about 528347193 people reported Hindi as their mother tongue.

As provided in the Constitution of India, the official languages of the Union are Hindi and English. Hindi is written in the Devanagari script, which follows a left-to-right direction and evolved from the ancient Brahmi script. Under the provisions of the Official Language Act, Hindi is employed in various governmental functions, including communication between the Central Government and those State Governments that recognize Hindi as an official language, as well as in parliamentary work and certain judicial activities.

1.4 HINDI PARALLEL CORPUS

Hindi is a moderately inflectional Indo-Aryan language with clear grammatical structure and rich linguistic features. It marks gender (masculine and feminine), number, and person agreement, and uses postpositions as case markers, with nouns occurring in direct and oblique forms. Verbs agree with the subject and, in certain perfective constructions, with the object. Compound as well as conjunct verbs are also widely used to express aspectual meanings. The normal word order is Subject–Object–Verb (SOV), though it allows some flexibility due to case marking and context. Hindi is written in the Devanagari script derived from the Brahmi tradition and written from left to right, while accurately representing phonemic distinctions such as aspirated–unaspirated and voiced–voiceless sounds along with short and long vowels. Linguistically, Hindi also shows features like ergative alignment in the perfective aspect, differential object marking, and productive compounding and derivation, while its vocabulary includes inherited Indo-Aryan words, Sanskrit borrowings, and loanwords from Persian, Arabic, and English, reflecting long historical contact and making it significant for linguistic study in India.

Hindi 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 HINDI PARALLEL CORPUS

The development of Hindi Parallel Corpus, English was used as the source language which was translated into Hindi. The Hindi 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, Dimasa, Dogri, Gangte, Garhwali, Garo, Gojri/Gujjari/Gujar, Gujarati, Gujar, Halabi, Handuri, Hara/Harauti, Haryanvi, Hindi Multani, 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, Kurmal 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 Hindi which are systematically structured based on 159 grammatical categories. Hindi has 32,905 word count and 147,098 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] LangLex. (n.d.). Hindi: Language profile (2011 Census report). Retrieved March 4, 2026, from <https://langlex.com/cens/LangProfile.php?lngname=HINDI>
- [2] Office of the Registrar General & Census Commissioner, India. (2018). C-16: Population by mother tongue (Table). Ministry of Home Affairs, Government of India. <https://language.census.gov.in/showMTSIdata>
- [3] Rejitha K. S. and Narayan Kumar Choudhary. (ed.). (2025). LDC-IL Corpus Insights. CIIL, Mysore. 978-93- 48633-33-0.
- [4] Verma, D. (1954). Hindi bhasha ka itihās [History of the Hindi language]. Hindustani Academy, Prayagraj.