



HARA/HARAUTI PARALLEL
TEXT CORPUS: LINGUISTIC
FEATURES AND STRUCTURES

HARA/HARAUTI 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. 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

Hara/Harauti Parallel Text Corpus: Linguistic Features and Structures

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

Contributors: Kishan Kumar, Atul Kumar Jain

e-ISBN: 978-81-69099-08-0

CIIL Publication No.: 1648

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

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

HARA/HARAUTI PARALLEL TEXT CORPUS

Dr. Satyaendra Kumar Awasthi, Teena Verma

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 HARA/HARAUTI: A BRIEF INTRODUCTION

Harauti is a rich and melodious dialect of Rajasthan, known for its heroic tales and unique grammar. It is a regional variety spoken primarily in the southeastern part of Rajasthan, known as the Hadoti region. The region's name is derived from the Hada branch of the Chauhan Rajputs, who had established their rule around 1241 AD. Following the defeat of Prithviraj Chauhan by Sultan Mohammed Ghauri, the Hadas migrated toward the Chambal valley under the leadership of Rao Dev. After enduring periods of Mughal vassalage and subsequent harassment by Maratha raiders, the local rulers sought British protection in 1818 to restore peace (Koshy, 2012). According to Greison, the Central Eastern dialect of Rajasthan has four speech varieties, namely, Jaipuri, Ajmeri, Krishnagarhi and Harauti. As per the census report of 2011, a total population of 29,44,356 reported their mother tongue to be Hara/Harauti. Hadothi is primarily spoken in the districts of Kota, Bundi, Jhalawar, and Baran.

Notably, the speaker population is concentrated entirely within Rajasthan, which stands in contrast to many other endangered languages, which often maintain fragmented speaker communities across multiple states. However, scholars observe that Hadoti is increasingly

threatened by prolonged contact with dominant languages like Hindi and English. This decline is driven by various sociolinguistic processes, including internal and external structural changes, extensive lexical borrowing, and significant functional and semantic shifts (Singh, 2018). Hadoti folk culture remains uniquely localized, with its traditional practices concentrated primarily within the districts of Kota, Bundi, Jhalawar, and Baran. This regional identity is defined by the renowned Hadoti School of Miniature Painting and celebrated through vibrant festivals like Kajli Teej. Furthermore, the area's heritage is preserved through rhythmic performance arts, such as the Gair dance, and a rich legacy of traditional handicrafts.

1.4 HARA/HARAUTI PARALLEL CORPUS

Harauti is a well-defined Indo-Aryan language characterised by a clear grammatical structure and distinct phonetic features. It is a culturally rich language with clear gender and number markings, verb aspect markings and a systematic use of word-order. Despite its close relationship to Hindi, and its shared borders with various Tibeto-Burman influences, it maintains its distinct identity through its independent form, phonetics, and popular usage.

Some features of Hara/Harauti are

- **Phonetic Properties:** One very distinct feature of Hara/Harauti is the process of reduplication. Sonority serves as a key prosodic feature in reduplication in Hara/Harauti. In this process, syllables typically organize around a high-sonority nucleus flanked by less sonorous onsets and codas. When the base or stem ends in an open syllable, the resulting reduplicant strictly adheres to the Sonority Sequencing Principle (SSP) (Chand & Kar, 2017).
- **Morphological Properties:** Nouns and pronouns exhibit distinct markers for number, gender, and case, while the verbal system is notably complex, utilizing aspectual markers and auxiliary verbs to denote tense and mood. Furthermore, the language frequently employs productive processes like reduplication and compounding to expand its lexical base and convey nuanced meanings (Koshy, 2012).
- **Syntactic Properties:** Sentences typically follow a Subject-Object-Verb (SOV) order, making it a verb-final language, a common feature across the Indo-Aryan Group.
- **Lexical Diversity:** The language's vocabulary shows visible influence from Hindi. Scholars report that this change or language assimilation happened over time. For example, the word for latch [sā:kʰə] is getting replaced by the Hindi word [kundi:] in general use. Similarly, the word for unripe tomato is [ko:kəl bəʈʈa:] but most people use the Hindi expression [kəʈʈʃa: ʈəma:ʈər].

Harauti 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 HARA/HARAUTI PARALLEL CORPUS

For the development of the Harauti Parallel Corpus, Hindi was used as the source language, which was translated into Harauti. The Harauti parallel text corpus is 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, 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, 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 Hara/Harauti which are systematically structured based on 159 grammatical categories. Harauti has 34,537 word count and 155,620 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] Chand, G., & Kar, S. (2017). Sonority and reduplication in Hadoti. *Journal of Universal Language*, 18(2), 1–37.
- [2] Koshy, B. (2012). Sociolinguistic survey of selected Rajasthani speech varieties of Rajasthan, India, volume 3: Hadothi (SIL Electronic Survey Report 2012-031). SIL International.
- [3] LangLex. (n.d.). Hara/Harauti: Mother tongue and its regions (2011 Census report). Retrieved March 5, 2026, from <https://langlex.com/cens/mtregion.php?mtname=Hara%2FHarauti>
- [4] Office of the Registrar General & Census Commissioner. (2011). *Census of India 2011: Language and Mother Tongue*. Government of India.
- [5] Rejitha K. S. and Narayan Kumar Choudhary. (ed.). 2025. LDC-IL Corpus Insights. Central Institute of Indian Languages, Mysore. 978-93-48633-33-0.
- [6] Singh, R. (2018). Revival of endangered languages: A case study of Hadoti. *International Journal of Research and Analytical Reviews (IJRAR)*, 5(4), 868–871.