SAURASHTRA/ SAURASHTRI

PARALLEL TEXT CORPUS:

LINGUISTIC FEATURES AND STRUCTURES

SAURASHTRA/SAURASHTRI
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:
Dr. Kamaraj S,
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

Saurashtra/Saurashtri Parallel Text Corpus: Linguistic Features and Structures

Authors: Dr. Kamaraj S, Dr. Rejitha K. S, Dr. Narayan Choudhary, Prof. Shailendra Mohan

Contributors: K. S. Senthilkumar, K S Malini

e-ISBN:
CIIL Publication No.: 1733

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: Dr.Kamaraj S

Table of Contents

SAURASHTRA/SAURASHTRI

PARALLEL TEXT CORPUS

Dr.Kamaraj S

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 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 SAURASHTRA/SAURASHTRI: A BRIEF INTRODUCTION

Saurashtra language (also known as Saurashtri) is an Indo-Aryan language predominantly spoken by the Saurashtrian community settled in Tamil Nadu, especially in cities such as Madurai, Thanjavur, and Salem. Smaller speaker populations are also found in Karnataka, Andhra Pradesh, and Kerala. According to the Census of India 2011, approximately 2.4 lakh people reported Saurashtra as their mother tongue. Saurashtra traces its origins to the Saurashtra region of present-day Gujarat. The ancestors of the present Saurashtrian community migrated to South India several centuries ago, likely between the 11th and 16th centuries, due to socio-political changes in western India. Although geographically separated from other Indo-Aryan languages, Saurashtra retained its Indo-Aryan grammatical base while absorbing significant lexical and structural influences from Dravidian languages such as Tamil and Telugu. It has also borrowed vocabulary from Sanskrit and, in modern times, from English. The language has its own distinct writing system known as the Saurashtra script. However, historically and even today, it has also been written in Tamil, Telugu, Devanagari, and Roman scripts depending on

regional and practical considerations. The Saurashtra script is phonologically systematic and capable of representing the sounds of the language, including aspirated consonants and vowel length distinctions typical of Indo-Aryan languages.

Saurashtra literature includes devotional songs, folk poetry, prose writings, translations, and modern creative works. Much of its earlier literary expression was transmitted orally, particularly through community songs and religious compositions. In the 19th and 20th centuries, efforts were made to standardize the script and promote written literature, resulting in printed books, magazines, and educational materials in Saurashtra. Saurashtra is a language with deep historical roots shaped by migration, cultural adaptation, and sustained community identity. Its unique position as an Indo-Aryan language thriving within a predominantly Dravidian linguistic region makes it particularly significant for linguistic and sociocultural studies. Though it has a relatively smaller speaker base compared to major Indian languages, it continues to flourish within the Saurashtrian community through cultural associations, print publications, and community-based educational initiatives.

1.4 SAURASHTRA/SAURASHTRI PARALLEL CORPUS

Saurashtra language (also known as Saurashtri) is an Indo-Aryan language with notable agglutinative tendencies and rich morphophonemic features, largely influenced by prolonged contact with Dravidian languages such as Tamil and Telugu. It exhibits morphophonemic alternations during suffixation, particularly in verb conjugation and nominal inflection. The language shows case marking, postpositional constructions, and productive compounding. Its word order generally follows the Subject–Object–Verb (SOV) pattern, though pragmatic factors allow some flexibility. Saurashtra possesses a unique script known as the Saurashtra script, which was developed to represent its phonological system. In addition to its own script, the language has historically been written in Tamil, Telugu, Devanagari, and Roman scripts, depending on the region and community practices. The Saurashtra script includes a well-developed set of consonant and vowel symbols capable of representing its phonemic distinctions, including aspirated stops and vowel length contrasts characteristic of Indo-Aryan languages. Structurally, Saurashtra reflects a blend of Indo-Aryan grammatical foundations and Dravidian areal features due to centuries of geographical and cultural interaction in South India. This combination makes it linguistically distinctive among Indian languages and significant for comparative linguistic research, particularly in the study of language contact, convergence, and structural adaptation.

Saurashtra is a valuable resource for developing parallel text corpora with other Indian mother tongues. A parallel corpus involving Saurashtra can provide important insights into syntactic divergence between Indo-Aryan and Dravidian languages, morphological complexity, code-mixing practices within multilingual communities, and script variation. Given that Saurashtra speakers are typically multilingual, the language offers rich data for examining cross-linguistic influence and translation strategies in multilingual contexts such as India.

This Saurashtra/Saurashtri parallel corpus constitutes a systematically translated dataset derived from an original English corpus, ensuring structured cross-linguistic alignment and semantic correspondence

1.5 SUMMARY OF SAURASHTRA / SAURASHTRI PARALLEL CORPUS

The Saurashtra/Saurashtri parallel text corpus is aligned with English and 146 mother tongues of India. These 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, 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, 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 Saurashtra/Saurashtri which are systematically structured based on 159 grammatical categories. Saurashtra/Saurashtri has 22666 word count and 153600 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.