

Parliament Question: Mother Tongue Based AI Learning System for Tribal and Multilingual Area

प्रविष्टि तिथि: 12 AUG 2026 4:49PM by PIB Delhi

The National Education Policy (NEP) 2020 highlights the importance of multilingualism and places strong emphasis on the promotion of all Indian languages. In alignment with the objectives of NEP 2020, the Government of India has undertaken several initiatives promoting education, preservation and research on Indian languages. These efforts have further been augmented through the adoption of Artificial Intelligence and Machine Learning (AI/ML) technologies, including the development of Large Language Models, translation tools and language-enabled digital services in Indian languages.

Government of India has initiated the BharatGen project, which is a multimodal Large Language Model (LLM) focused on developing efficient and inclusive AI solutions that support all 22 scheduled languages and enable the creation of a robust digital AI infrastructure. BharatGen is anchored at IIT Bombay, where the core development of language technologies is undertaken. The initiative is implemented through a consortium of leading academic institutions including IIT Kanpur, IIT Madras, IIT Hyderabad, IIIT Hyderabad, IIM Indore, and IIT Mandi.

Further, to facilitate the translation of content into Indian languages, technological advancements include the development of AI-based translation tools such as ANUVADINI by the All-India Council of Technical Education (AICTE) and BHASHINI, an initiative under the Digital India Programme.

BHASHINI has developed state-of-the-art AI models for Indian languages through a collaboration of over 70 research partner institutes. BHASHINI platform hosts a repository of over 360 AI-based language models and provides more than 22 specialized language services. These services include Automatic Speech Recognition (ASR), Machine Translation (MT), Text-to-Speech (TTS), Optical Character Recognition (OCR) and Transliteration. The dataset corpus includes 246 million parallel sentence pairs and 3.7 million monolingual text entries. All datasets and models are publicly accessible via BHASHINI platform or through the Digital India BHASHINI Division account on the AIKosh platform. Further, the National Language Translation Mission through the BHASHINI platform is digitising large volumes of text and speech data across all 22 scheduled languages. The platform also supports tribal languages such as Bhili and Santhali.

In addition, the Central Institute of Indian Languages (CIIL), a subordinate office under the Ministry of Education in collaboration with NCERT, has developed foundational primers and linguistic datasets in various tribal languages including Santhali. Furthermore, it has launched a 12-week Santhali language course on the SWAYAM platform to promote mother tongue-based education.

Under the Linguistic Data Consortium for Indian Languages (LDC-IL), a project of CIIL, linguistic resources have been developed for tribal languages, including Santhali and Mundari. The resources available for Santhali include a speech dataset comprising 50 hours of audio recordings from 60 speakers. In addition, parallel text corpora covering linguistic features and structures have been developed, comprising 5,200 sentences each in Santhali and Mundari. LDC-IL has also developed web-

based applications and tools to support research and development in Indian languages. Several of these tools, hosted at CIIL data centres, are accessible through the Medha Bhashika website <https://medha.ciil.org>

CIIL has also prepared foundational primers in several languages namely Mundari, Kharia, Ho-Hindi, Kurukh, Korwa (Jharkhand and Chhattisgarh), Bhumij, Santhali-Hindi and Malto. Further, CIIL's Bharatavani portal hosts diverse Santhali resources such as Bhashakosha (language learning repository), Jnanakosha (encyclopaedia), Pathyapustakakosha (textbooks) and Shabdakosha (dictionaries).

This information was given by **the Minister of State for Education, Dr Sukanta Majumdar** in a written reply in the Rajya Sabha today.

RRTN

(रिलीज़ आईडी: 2298335) आगंतुक पटल : 497

इस विज्ञप्ति को इन भाषाओं में पढ़ें: Urdu , हिन्दी , Bengali