GOVERNMENT OF INDIA
MINISTRY OF ELECTRONICS AND INFORMATION TECHNOLOGY
LOK SABHA
UNSTARRED QUESTION NO. 486
TO BE ANSWERED ON: 03.12.2025

TAMIL LANGUAGE DATASETS

486. THIRU DAYANIDHI MARAN:

Will the Minister of ELECTRONICS AND INFORMATION TECHNOLOGY be pleased to state:

(a) whether the Government has created or funded Indian-language AI datasets under the India AI Mission for National Language Processing (NLP) speech recognition, machine translation and content moderation and if so, the status and availability of such datasets and the funds allocated for the same itemised language-wise;
(b) the funds allocated and actually utilised for developing these datasets during each of the last five years by language-wise including the details of implementing agencies;
(c) whether the Government proposes to create dedicated, high-quality datasets in Tamil including spoken variants to improve AI accuracy, safety and moderation in Tamil digital ecosystems and if so, the details specific steps taken, timelines and budgetary support provided thereof;
(d) whether any consultations have been held with Tamil linguistic experts, universities or research institutions to guide the development of such datasets and if so, the details thereof; and
(e) the total amount of funds budgeted, allocated, sanctioned and spent on tamil language datasets?

ANSWER

MINISTER OF STATE FOR ELECTRONICS AND INFORMATION TECHNOLOGY
(SHRI JITIN PRASADA)

(a) to (e): India’s AI strategy is based on the Hon’ble Prime Minister’s vision to democratize the use of technology. It aims to address India centric challenges and create opportunities for all Indians.

IndiaAI Mission:
It is a strategic initiative to establish a robust and inclusive Al ecosystem aligned with India's development goals through the seven pillars — IndiaAI Compute, AIKosh, IndiaAI Foundation Models, IndiaAI FutureSkiIIs, Startup Financing, Application Development and Safe & Trusted Al.

Around 275 datasets have been uploaded on AIKosh. This includes 100+ Tamil language datasets contributed by leading organisations.

The IndiaAI Mission supports the creation and funding of Indian-language Al datasets for various use cases including NLP, speech, translation and content moderation.

Mission BHASHINI:

Mission BHASHINI under the Digital India Programme enables multilingual digital access by developing Al-powered speech and text technologies for Indian languages. It functions as the official repository under the National Language Translation Mission (NLTM).

BHASHINI has been developed through a national collaboration of over 70 research partner institutes. It hosts a repository of over **350 AI-based language models** and offers **22+ specialized language services.** It offers Automatic Speech Recognition (ASR), Machine Translation (MT), Text-to-Speech (TTS), Optical Character Recognition (OCR), and Transliteration.

.

Dataset corpus includes:

- Ø 246 million parallel sentence pairs
- Ø 3.7 million monolingual text entries
- Ø 14,000 hours of ASR audio
- Ø 2.5 million OCR image samples
- Ø 476 hours of TTS audio
- Ø 20.56 million transliteration entries across Indic languages

All datasets and models are publicly accessible via the BHASHINI platform or through the Digital India BHASHINI Division account on the AIKosh platform.

Dedicated datasets for Indian languages including Tamil and spoken variants have been created. For ASR, transcribed speech datasets and trained models are available for 22 scheduled Indian languages, including Tamil, contributed by A14Bharat at IIT Madras.

The Tamil datasets include parallel and monolingual corpora, ASR (labelled/unlabelled), TTS, OCR, transliteration, terminology, **NER (Named Entity Recognition)** and glossary resources, publicly accessible through the BHASHINI and AIKosh platforms.

Development work has been carried out with linguistic experts, research and academic partners, through structured consultation.

A total fund of INR 47 crore was allocated for building datasets for 22 scheduled Indian languages across Translation, Speech Recognition, Speech Synthesis and OCR activities, and the entire amount has been fully utilised.

*****