Technology Development for Indian Languages

India is a multilingual country, with 22 official languages and 12 scripts. In India only about 5% - 10% people know English and the rest are deprived of benefits of advances in information technology. The benefits of information technology can reach to the common man only when software tools and human machine interfaces are available in people's own languages. Technology Development for Indian Languages (TDIL) Programme of Ministry of Electronics & Information Technology (MeitY), Government of India is an on-going Research Programme with following vision & mission objectives:

Vision:

- Digital Unite & Knowledge for All.

Mission:

- Proliferation of Language Technology
- Research and Development of Language Technology
- Development of Standards related to Language Technology

Outcome of TDIL Programme:

The research efforts undertaken through TDIL programme of MeitY have lead to following outcomes with regard to proliferation of Indian Languages:

(A) Proliferation

Free Language CD:

• It contain software tools for 22 Indian Languages (for both Windows and Linux Operating Systems) have been made available for download from http://ildc.in.

•Total Language CD dispatched – 13 Lakh and total Language CDs downloaded are 1.58 Crore.

•Tools like Localized Libre Office, Unicode Typing Tool, Unicode fonts, Localized Firefox Browser and many more are freely made available.

Virtual keyboard for 22 Indian languages which can be downloaded for use on mobiles has been made available on https://apps.mgov.gov.in has proliferated use of Indian languages on mobile devices.

(B) Research and Development

Machine Translation Systems:

• Indian Language to Indian Language Machine translation system in 18 language pair
• English to Indian Language MT (AnglaMT) for 8 language pairs.
• English to Indian Language MT (Anuvadaksh) for 8 language pairs.
• Hindi to English Language MT (HEMAT) for Judicial domain.

Text To Speech (TTS) system for 10 Indian languages viz. Tamil, Telugu, Marathi, Bodo, Kannada, Odia, Hindi, Malayalam, Manipuri & Rajasthani.

Automatic Speech Recognition (ASR): Automatic Speech Recognition (ASR) for Agricultural Commodity prices and weather information system in 11 Indian Languages/Dialects have been developed, namely Tamil, Telugu, Marathi, Hindi, Bengali, Assamese, Hindi (Bihari), Hindi (Jharkandi), Kannada, Gujarati, Odia. The systems would act as a voice interface for NIC Agmarknet portal.

Optical Character Recognition in Indian Languages:

Optical Character Recognition engine for 10 Indian languages namely Bengali, Devanagari, Gurumukhi, Kannada, Malayalam, Telugu, Tamil, Assamese, Odia and Urdu languages. Web OCR (for 10 ILs) as well as Desktop OCR (for 7 ILs) is made available for public use.

On-line Handwriting recognition system (OHWR):

Online Handwriting Recognition in 8 Indian Languages namely Hindi, Bengali, Tamil, Telugu, Kannada, Malayalam, Assamese and Punjabi languages have been developed.

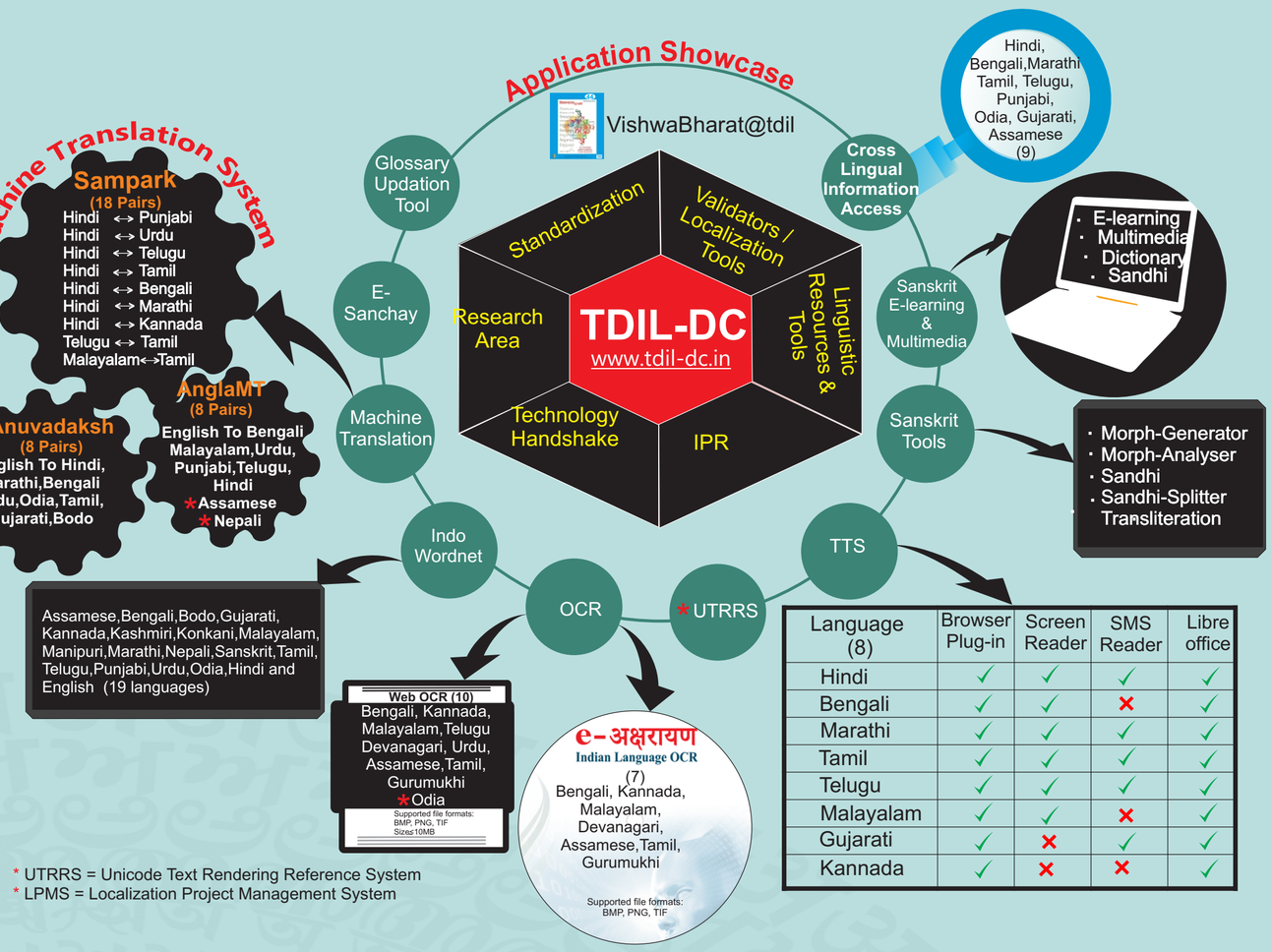
Cross Lingual Information Access (CLIA):

Monolingual Search Engines for Tourism Domain for 9 Indian Languages (Sandhan) –Hindi, Bengali, Marathi, Tamil, Telugu, Punjabi, Odia, Gujarati and Assamese.

Linguistic Resources:

Resources like wordnet, text corpora, speech corpora , Treebank text data, tts voices, ASR data have also been developed to support and promote research in Indian Language.

"Dissemination Through http://tdil-dc.in "



Language (8)	Browser Plug-in	Screen Reader	SMS Reader	Libre office
Hindi	✓	✓	✓	✓
Bengali	✓	✓	×	✓
Marathi	✓	✓	✓	✓
Tamil	✓	✓	✓	✓
Telugu	✓	✓	✓	✓
Malayalam	✓	✓	×	✓
Gujarati	✓	×	✓	✓
Kannada	✓	×	×	✓

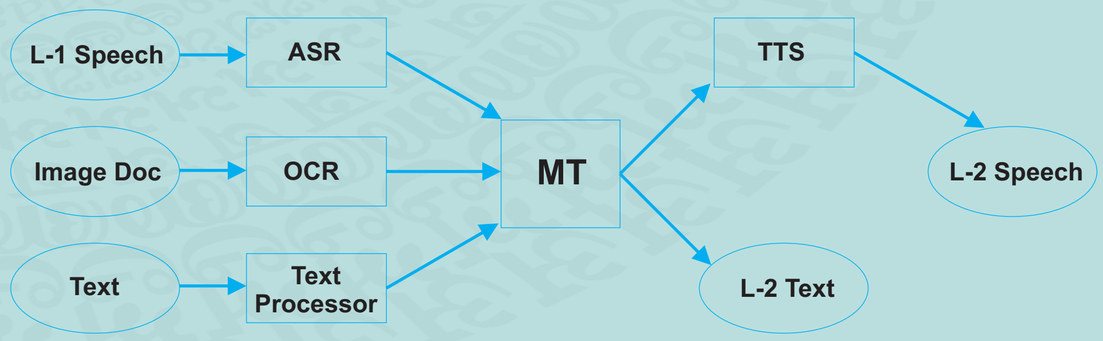
* UTRRS = Unicode Text Rendering Reference System
* LPMS = Localization Project Management System

(C) Standardization :

Efforts have been taken in standardization of Indian language in Unicode, Standardize Inscript keyboard layout. Also worked with BIS and ICA for mandating support of Hindi on all mobile devices.

New Initiative: Natural Language Translation Mission

Mission is to overcome the language barrier for all major languages of India in speech and text form.



SSMT System Flow

Deliverables:

(a) Speech to Speech Machine Translation (SSMT) system for major Indian languages: It would be possible with minimal human involvement but the involvement would decrease with the time as the machine learns on its own. Adaptation of the system in the domains within broad areas of science & technology, education, healthcare, governance, law & justice, etc.

(b) Text-to-Text Machine Translation system for major Indian languages: As in the case of SSMT system, it would be possible with minimal human involvement but the involvement would decrease with the time as the machine learns on its own. Adaptation of the system in the domains within broad areas of science & technology, education, healthcare, governance, law & justice, etc.

(c) Deployment of language technology-based applications through more than 100 start-ups.

(d) National Platform for Language Technology – offerings of linguistic resources and language technology-based services at competitive cost which will decrease progressively.

(e) Five-fold increase in content in Indian languages on Internet .

(f) A portal for showcasing technology and carrying out assessment of technology.

(g) Hackathons and grand challenges in the area of language technology .

(h) Repository of standards, best practices and benchmarks for Indian languages.