Situation and Challenges of Technologies for Indigenous Languages of India

Shweta Sinha[1]
[1]Amity University, Gurgaon
meshweta_7@rediffmail.com

Shyam Sunder Agrawal[2]
[2]KIIT College of Engineering, Gurgaon
ss_agrawal@hotmail.com

Introduction

India is a country with huge linguistic diversity. Out of 900 languages spoken in the country, only a few have witnessed the digital world. This poster presents in detail the Indian languages situation in terms of resources, and technologies. It highlights the relative needs, opportunities, barriers and complexities specific to the Indian Languages technologies. The aim is to study their influence on the adoption and adaptation of digital technology vis a vis Technological achievements/ fallout's of Indian languages relating to the world languages and to identify the gap and the need to take up future projects for technological advancements.

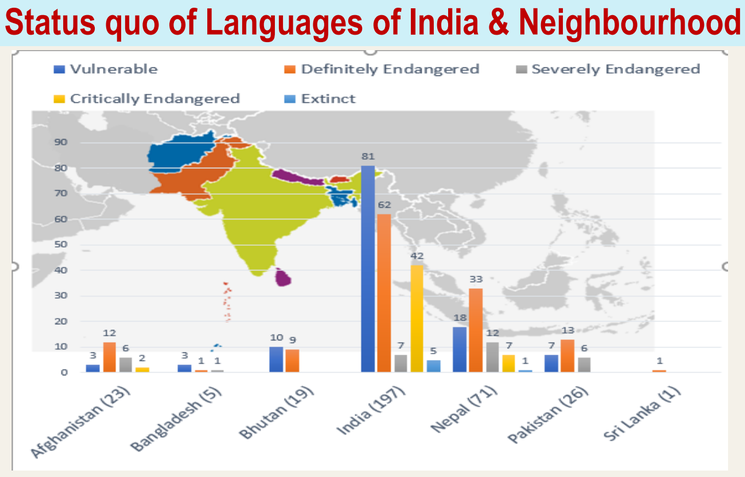
भारत एक विशाल भाषाई विविधता वाला देश है। देश में बोली जाने वाली 900 भाषाओं में से कुछ ही डिजिटल दुनिया में देखी गई हैं। यह पोस्टर संसाधनों, और प्रौद्योगिकियों के संदर्भ में भारतीय भाषाओं की स्थिति को विस्तार से प्रस्तुत करता है। यह भारतीय भाषाओं की प्रौद्योगिकियों की विशिष्ट आवश्यकताओं, अवसरों, बाधाओं और जटिलताओं को उजागर करता है। इसका उद्देश्य डिजिटल प्रौद्योगिकी को अपनाना और उनके अनुकूलन पर उनके प्रभाव का अध्ययन करना है। विश्व भाषाओं से संबंधित भारतीय भाषाओं की तकनीकी उपलब्धियों नतीजों और अंतर की पहचान करना और भविष्य की परियोजनाओं को पूरा करने की आवश्यकता है।

Language Diversity in India

- Languages in India (Census –GOI) : 122 languages and 2371 dialects
- Spoken Languages in India (PLSI) : 780
- Languages lost in last 50 years : 250
- Constitutionally recognized Official Languages :22

Language	2011 Census of India[18] (total population 1,210,854,977)[19]	Language	2011 Census of India[18] (total population 1,210,854,977)[19]
	Percentage		Percentage
Hindi[20]	43.63%	Gondi	0.25%
Bengali	8.03%	Nepali	0.25%
Marathi	6.86%	Sindhi	0.24%
Telugu	6.70%	Dogri	0.22%
Tamil	5.70%	Konkani	0.19%
Gujarati	4.58%	Kurukh	0.16%
Urdu	4.19%	Khandeshi	0.15%
Kannada	3.61%	Tulu	0.15%
Odia	3.10%	Meitei (Manipuri)	0.15%
Malayalam	2.88%	Bodo	0.13%
Punjabi	2.74%	Khasi	0.12%
Assamese	1.26%	Ho	0.12%
Maithili	1.12%	Mundari	0.09%
Bhili/Bhilodi	0.86%	Garo	0.09%
Santali	0.65%	Tripuri	0.08%
Kashmiri	0.58%		

Status quo of Languages of India & Neighbourhood



Languages Covered for Technology Development

Speech Database	ASR	TTS
Hindi	Hindi	Hindi
Tamil	Tamil	Tamil
Bengali	Bengali	Bengali
Telugu	Telugu	Telugu
Marathi	Marathi	Marathi
Indian English	Indian English	Malayalam
Assamese	Assamese	Assamese
Punjabi	Punjabi	Kannada
Manipuri	Manipuri	Manipuri
Bodo	Bodo	Bodo
Gujarati	Gujarati	Gujarati

Languages Resources Developed

Text Resources Available

S No	Languages	Corpus Detail	Statistics	Institution
1.	Malayalam	Malayalam Treebank Data	9512 monolingual Sentences, 6010 parallel sentences	IIIT Hyderabad
2.	Kannada	Kannada Treebank Data	19550 monolingual Sentences	IIIT Hyderabad
3.	Marathi	Marathi Treebank Data	10852 sentence , 3450 parallel sent	
4.	Hindi	Hindi Treebank Data Hindi Monolingual Text ILCI II	3000 Sentences	IIIT Hyderabad JNU
5.	Urdu,Tamil, Punjabi, Odia, Bodo, Bangla Assamese	Monolingual Data		JNU
6.	AngleMT	Nepali, Malyalam(Health) etc	23000 Malyalam Sentences	CDAC

Speech Recourses Developed

Sl No	Resources	Language & Statistics	Organization
1	PLS	Hindi:50,000 lexemes ,Marathi : 51,065 lexemes Punjabi: 33,874 lexemes, Manipuri : 2,83,998 lexemes Assamese: 53,304 lexemes	TDIL
2	Speech samples in agriculture domain	Telugu 1073 speakers, Tamil 1000 speakers Marathi 1500 speakers, Bangla 1000 speakers Assamese 1023 speakers	TDIL
3	Annotated speech samples	Bengali 450 speakers, Hindi 650 speakers Konkani 450 speakers, Odia 450 speakers Malayalam & Tamil 450 speakers	LDC-IL
4	GlobalPhone	2000 native speakers transcribed data in Tamil	ELRA
5	EMILLE/CIIL Corpus	Monolingual, parallel and annotated corpora in Assamese, Bengali, Gujarati, Hindi, Kannada, Kashmiri, Malayalam	ELRA
6	Annotated Speech Samples	Assamese: 5658 speech data files; 27 speakers Bengali :2500 speech data files; 21 speakers Nepali: 660 speech files; 6 speakers English :2500 speech files; 16 speakers	IIT Guwahati
7	Prosody model development	Gujrati:1000 speakers IVR recording, Audio search system, ASR Marathi: 1000 speakers IVR recording,Audio search system, ASR	DAICT, Gandhi Nagar
8	Prosodic word Dictionary	English: 5031 word dictionary generated from 2500 spoken Bengali sentences	IIT Kharagpur

Technology Developments

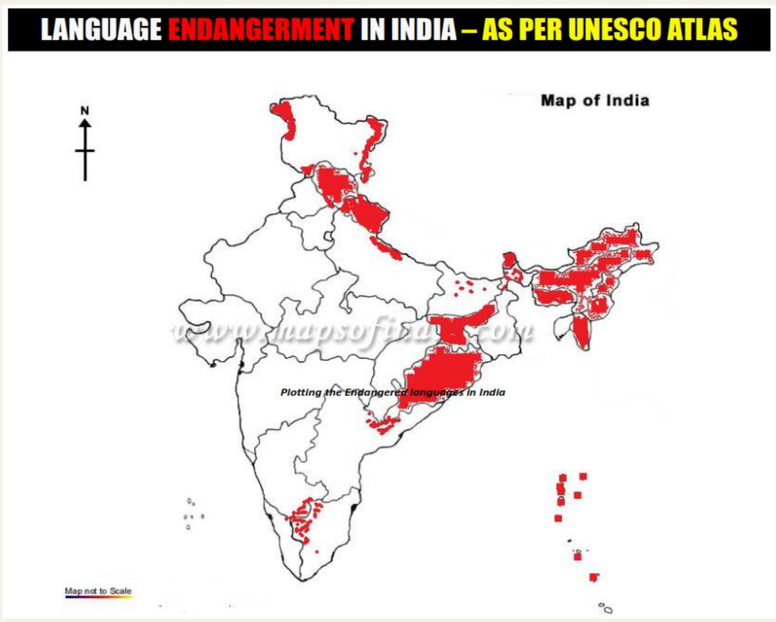
Document Analysis and Recognition (OCR Technologies)

e-Aksharayan: A Desktop software for converting scanned printed Indian Language documents into a fully editable text format in Unicode encoding. Works on Windows 7,8, and 10. Equipped with Unicode typing tool for typing in Indian Language and Sakal Bharati font (11 Indian Language scripts in a Single font).

Available for Assamese, Bangla, Punjai, Hindi, Kannada, Malayam, Tamil languages

Features:

- **Recognition :** it enables users to harness the power of computers to access printed documents in Indian language/scripts.
- **Availability:** Present version of e-Aksharayan supports major Indian languages- Hindi, Bangla, Malayalam, Gurmukhi, Tamil, Kannada & Assamese.
- **Accuracy:** It converts printed document images to editable text with upto 90-95% recognition accuracy at character level & 85-90% at word level.
- **Speed:**Current version of e-Aksharayan takes 45 to 60 sec to process an A4 size document.



Machine Translation

System	Approach used	Target Language	Place	Features
Hindi to Punjabi MTS	Direct	Hindi to Punjabi	Punjab Patiala Uniiversity	Morphological analysis, word sense disambiguation, post processing and transliteration
Mantra	Transfer Based	English to Hindi, Gujrati, Telegu. Hindi to English, Bengali, Marathi	CDAC,Pune	Uses Tree Adjoining Grammar Formalism.
Anubaad	Transfer Based	English to Bengali	CDAC, Kolkata	Hybrid system which uses n-gram approach for POS tagging. Works at sentence level
Anglabharti	Interlingual	English to Hindi, Tamil	IIT,Kanpur	Uses intermediate structure Pseudo Lingua for IL.
AnglaHindi	Interlingual	English to Hindi	IIT,Kanpur	Uses rule-bases, example-base and statistics.
English Hindi MTS	Statistical Machine Translation	English to Hindi	IIIT, Hyderabad	Combines Rule Based Machine Translation and phrase based SMT
English Malyalam MTS	Statistical Machine Translation	English to Malyalam	Cochin University	monolingual Malayalam corpus and a bilingual English/ Malayalam corpus in the training phase
Hindi-English MTS	Statistical Machine Translation	English to Hindi	State University of New York and IIT Kanpur	Combines Rule Based Machine Translation and phrase based SMT
Anubharti	Example Based Machine Translation	Hindi-English	IIT, Kanpur	Hybrid Example based system which combines pattern based and example based approach.

Text to Speech System

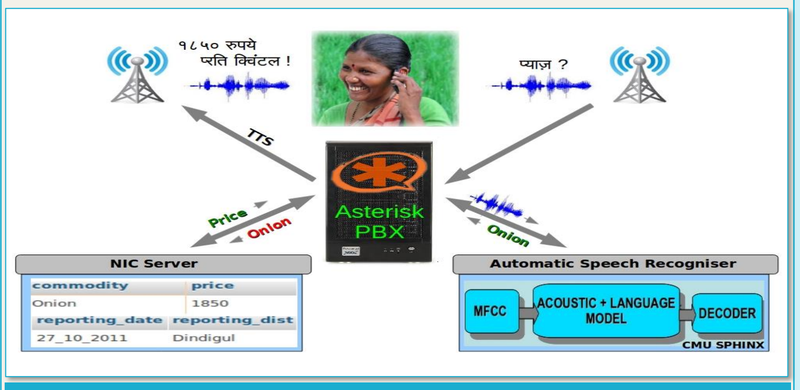
Sl No	Name of the Language	Development of TTS Engines (Concatenative and Statistical approach)
1.	Hindi	Male and Female USS and HTS, Male Bilingual HTS, Male
2.	Assamese	Male and Female voice – HTS, USS
3.	Bengali	Male and Female – HTS, USS
4.	Gujarati	Male HTS, USS, Male HTS using STRAIGHT approach.
5.	Marathi	Male and Female – HTS, USS, Male HTS STRAIGHT
6.	Malayalam	Male and Female – HTS, USS
7.	Kannada	Male HTS
8.	Manipuri	Female – HTS
9.	Tamil	Male and Female, Bilingual – HTS, USS
10.	Telugu	Male, Female – USS, HTS, Male HTS STRAIGHT
11.	Rajasthani	System under development
12.	Bodo	System under development

TTS Applications Developed

- **Browser Plug-in for Mozilla and Chrome Browser for Eight Indian Languages :** Hindi, Marathi, Bengali , Assamese, Tamil , Telugu , Malayalam, Kannada, Odia . Available through TDIL Data Centre portal : http://tdil-dc.in
- **SMS Reader for Android Platform :**The SMS reader (sandesh Pathak) has been made available in public domain free of cost through mobile-seva gateway http://mgov.gov.in
- **Screen Reader for Visually Challenged persons**

ASR in Indian Languages

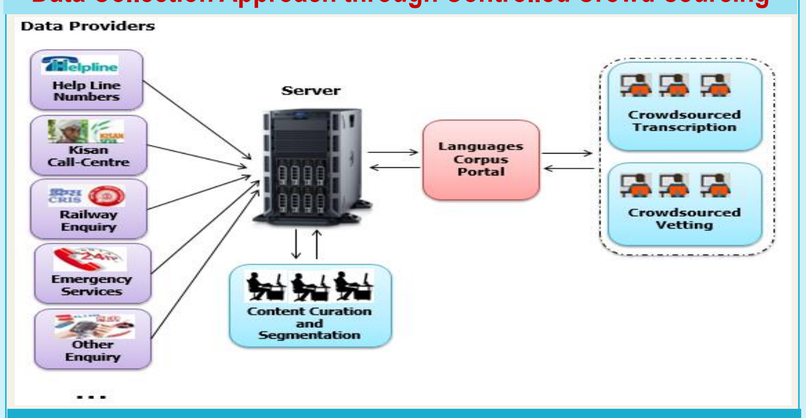
- **Automatic Speech Recognition for Agricultural Commodity prices for 6 Indian Languages developed namely in Hindi, Bengali, Assamese, Tamil , Telugu and Marathi**
- **Deployment of the ASR systems in collaboration with Ministry of Agriculture has been initiated for Tamil, Telugu, Bangla, Assamese and Marathi languages.**
- **Voice based Internet Browsing System in Hindi for Health Domains**



Challenges in Technology Development for IL

- **Resource creation based on global standards**
- **Language ambiguity and complexity: one word different meaning in different context**
- **Origin of Indian script and family: one language – many script; many language – one script**
- **Lack of language grammar, literature and documents standardization**
- **Difficulty in data collection due to geographical and social and cultural strata**
- **Presence of several dialects: code-mixing between dialects; massive number of non-native speakers**
- **Non-conformance with English centric models: existing models can't be extended to IL**
- **Localization issues due to operating system and applications**
- **Lack of encoding standards for several phones of Indian languages**

Data Collection Approach through Controlled Crowd-sourcing



Way Forward for IL Technology Development

- **Producing a White paper : reflect situation for all languages**
- **Massive amount of text data creation to reliably train statistical language model : phonetically balanced data**
- **Transcribed recordings from several speakers to obtain varying acoustical characteristics due to nativity and other socio linguistics aspect for creation of acoustic model**
- **Pronunciation dictionary of the vocabulary for lexical/PLS development: capturing prosody**
- **Urgency to work with zero resource language: avoid its extinction**
- **Availability of facilities such as BLARK (basic language resource tool) for all IL**