Proceedings of the Language Technologies for All (LT4All) , pages 324–327
Paris, UNESCO Headquarters, 5-6 December, 2019.

Situation and Challenges of Technologies for Indigenous Languages of India

S Sinha, S S Agrawal
Amity University Haryana, India[1], KIIT College of Engineering, Gurgaon, India[2]
ssinha@ggn.amity.edu[1], ss_agrawal@hotmail.com[2]

Abstract

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 paper presents the language situation in India. It also highlights the opportunities, barriers and complexities faced by the language technology community in the development of indigenous Indian languages. The aim is to study their influence on the adoption and adaptation of digital technology. Technological achievements/fallouts of Indian languages relating to the world languages will be analysed with the purpose to identify the gap. The paper also outlines the need for future language resources and uptake of projects for technological advancements of indigenous languages of India.

Keywords: Language Technology, Indian languages, Under-resourced Language

Résumé

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 पेपर 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 उपलब्धियांऔर अंतर की पहचान कर भविष्य की परियोजनाओं को पूरा करने की आवश्यकता है।

1. Introduction

India is a plurilingual and pluri-ethnic land. Linguistic diversity and multilingualism are essential for the enrichment of humanity and development and language is an important attribute of it. According to the census (2011), there are 121 languages and around 2300 dialects in India. These languages belong to five language family(census, 2011): the Indo-European (Indo-Aryan 78.05%), Dravidian (19.64%), Austro-Asiatic, Tibeto-Burmese and Semito-Hamitic. Out of all the languages spoken in the country, 22 languages are constitutionally recognized and 'Hindi' has the status of the official and national language (Jha, 2010). Table 1 presents the language and speaker population of major languages of the country.

In India, there are around 30 languages with more than one million population, but most of them have not seen the light of the digital world. This situation puts the users of indigenous languages in a disadvantageous situation. It creates a digital divide among the languages and puts them in danger of digital extinction that may lead to complete extinction also. The development of language technologies provides opportunities to exchange ideas with one another easily. Research community working in the area of language technology look forward to utilizing technological growth to create a workable platform. They aim to cater to the need of users irrespective to their language, age, gender and socio-economic background.

Undoubtedly, long term effort is required to cover all the languages and take benefit from digital growth. This paper highlights the language situation and technological growth for Indian languages. The analysis of the current situation helps to identify the existing challenges and barriers for language users. In the end, the paper outlines the need for future language resources and uptake of projects for technological advancements of indigenous languages of India.

Languages	Population(%)	Languages	Population(%)
Hindi	43.63	Malayalam	2.88
Bengali	8.03	Punjabi	2.74
Marathi	6.86	Assamese	1.26
Telugu	6.70	Maithili	1.12
Tamil	5.70	Santhali	0.61
Gujrati	4.58	Kashmere	0.56
Urdu	4.19	Nepali	0.24
Kannada	3.16	Sindhi	0.23
Odia	3.10	Dogri	0.21

Table 1: Major languages and speaker population of India

2. Digital Representation of Indian Languages

The three Indian Languages; Hindi, Punjabi and Bangla are among the top ten most widely spoken languages of the world (Arora et al.,2013), but none of these finds their place in the top ten languages on the web (Sinha et al.,2018). According to Unesco's "Atlas of the world's languages in danger " (Language atlas, 2009). India has the maximum number of endangered languages and most of the Indian languages are vulnerable. Availability of digital data for the languages may help in their revival. Online services in these languages may increase their user base. Limited language support and content are the largest barriers to the adoption of online services.

The internet contents are majorly available in the languages of developed countries with English having the topmost share with 56%, followed by Russian and Spanish with

7.3% and 4.7% respectively (Arora et al., 2013). Only 0.1% of the web content is available in Hindi language and none of the other Indian languages finds its place in the top 40 languages of the digital world. As far as language technology is concerned Hindi, Bangla, Telugu and Tamil are a few Indian languages that have some associated language technology with various quality level. Lack of digital resources for the Indian languages categorize them as under-resourced languages. Technology development is essentially required for keeping these languages alive. Efforts have been made to provide technical support to the Indian languages, but lack of resources makes it a challenging task. To start with, the researchers have worked for the technology development for few languages as mentioned above and is in the process of generating resources for many more languages.

3. Technology Development for Indian Languages

Resource creation is the first and the foremost important step towards the technology development for languages. Indian languages falling in the category of under-resourced languages require special effort for corpus creation. Efforts made in this direction have helped the researchers to create corpora for a few Indian languages and application based on ASR, TTS and MT have been developed for a few languages.

3.1 Development of Language Resources

Language technology is a data-centric research area. Text and speech data of any language are the necessity for developing technology for that language. Several research groups, the Ministry of Human Resource Development (MHRD) and its agency for language development; CIIL (Central Institute of Indian Languages) and Ministry of Electronics & Information Technology (MeitY) with its agency TDIL (Technology Development for Indian Languages) is continuously putting efforts for developing Indian languages resources.

3.1.1 Text Corpus Availability

Collection of phonetically rich text sentences are available for Malayalam, Kannada, Marathi, Hindi, Tamil, Punjabi, Bangla, Indian English and Assamese. The size of these corpus ranges from 3000 sentences to around 23000 sentences. Only a few of these databases are multilingual, and most of it is monolingual.

3.1.2 Speech Corpus for Indian Languages

The development of speech corpus is a time and labour intensive task. The created corpus has to be annotated before being utilized for technology development. The corpus collection was initially done in the studio environment with the aim to reduce noise and external interferences. Studio recording posses restriction in mobility. Since most of the Indian languages are under-resourced so availability of speakers for studio recording is difficult. With the advancements of technology, availability of noise reduction techniques and to match with the real-time scenario, studio recordings are not being used nowadays. The demand for corpus created in an office environment or noisy areas like roadside, moving vehicle or market place prevails now. Efforts are continued to collect resources using crowdsourcing (Arora et al., 2016) or through online platforms (Sinha et al.,2017) as a read or spontaneous speech. Table 2 presents the statistics of speech resources developed and reported so far.

Resources	Language & Statistics	Organization
PLS	Hindi: 50,000 lexemes, Marathi: 51,065 lexemes, Punjabi: 33,874 lexemes, Manipuri: 2,83,998 lexemes Assamese: 53,304 lexemes	TDIL
Speech samples: agriculture domain	Telugu 1073 speakers, Tamil 1000 speakers, Marathi 1500 speakers, Bangla 1000 speakers, Assamese 1023 speakers	TDIL
Annotated speech samples	Bengali 450 speakers, Hindi 650 speakers, Konkani 450 speakers, Odia 450 speakers, Malayalam & Tamil 450 speakers	LDC-IL
Global Phone	2000 native speakers transcribed data in Tamil	ELRA
EMILLE/ CIIL Corpus	Monolingual, parallel and annotated corpora in Assamese, Bengali, Gujarati, Hindi, Kannada, Kashmiri, Malayalam	ELRA
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
Prosody model develop-ment	Gujrati:1000 speakers IVR recording, Audio search system, ASR Marathi: 1000 speakers IVR recording, Audio search system, ASR	DAICT, Gandhi Nagar
Prosodic word Dictionary	English: 5031-word dictionary generated from 2500 spoken Bengali sentences	IIT Kharagpur

Table 2: Speech resources available for Indian languages

3.2 Development of Language Technology

Every human being wishes to deliver and also obtain information and services in their language. Today, for linguistic preservation and cultural redemption, development of language technology and digital representation of languages has become essential. Application based on automatic speech recognition, text to speech synthesis and machine translation makes life easier for people who like to avail facilities in their native language. Some efforts in this direction have been made for a limited number of Indian languages.

3.2.1 Automatic Speech Recognition (ASR)

Literature (Singh et al.,2019) highlights that the language research community of India has carried out several experiments for different Indian Languages based on small databases collected for experimental purpose. Most of these are done as a laboratory experiment and are not

converted into applications for general use. Researchers and industry have majorly focussed on HMM and ANN based methodologies for the development of ASR systems. A big giant like Google has created Assistant that converse in 6 Indian languages apart from many world languages. But, again this is a very small fraction as compared to 121 languages. Several prototypes such as railway enquiry system (Samudravijaya,2000), Bangla digit recognizer, travel enquiry system etc. have been developed by Indian research institutes, A major breakthrough in this direction is achieved by the development of a system for agriculture commodity prices. Speech-based access for Agricultural Commodity prices for 6 Indian Languages was developed for Hindi, Bengali, Assamese, Tamil, Telugu and Marathi. The project was carried as a consortium project supported by DeitY, India. The system uses an HMM-based speech recognizer and is helpful to illiterate farmers and visually impaired people. But, again this type of system lacks in catering to the dialectal, prosodic and tonal variations.

3.2.2 Text to Speech Synthesis (T-T-S)

TTS system when integrated with a screen reader is potentially assistive technology for visually impaired people. Concatenative and statistical approaches have been used to develop TTS engine for some of the Indian languages. TTS applications developed so far for Indian languages are as mentioned below(TDIL):

- **TTS integrated with Screen Reader for Visually Challenged persons:** TTS integrated with Screen Reader are available in Hindi, Bengali, Marathi, Tamil, Telugu and Malayalam.
- **Browser Plug-in**: TTS as browser plug-ins are also developed for eight Indian Languages namely Hindi, Bengali, Marathi, Tamil, Telugu, Malayalam, Odia and Gujarati.
- **SMS Reader in Indian Languages - Sandesh Pathak**: SMS Reader is an Android App and is made available for 5 Indian Languages namely Hindi, Marathi, Tamil, Telugu and Gujarati. Click here to download

These projects were carried out in consortium mode under the leadership of IIT Madras. Apart from this few works in this direction for other languages have also been reported. Table 3 presents the details.

Sl No	Name of the Language	Development of TTS Engines (Concatenative and Statistical approach)
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

Table 3: Details of TTS system development efforts

3.2.3 Machine Translation

In a multilingual country like India, a huge amount of information exchange takes place across different languages inside the country. It is thus necessary to have an automated process to convert data from the sender's language to that of the receiver's language. Efforts have been put in this direction to obtain automatic translator of languages. Most of the work till date is confined between Hindi and English language pair but other language pair are also being tried now. Table 4 presents some of the well known MT systems available in Indian languages. The approach used by them ranges from example-based to rule-based and to statistical MT systems.

System	Target Language	Place	Features
Mantra	English to Hindi, Gujrati, Telegu. Hindi to English, Bengali, Marathi	CDAC,Pune	Uses Tree Adjoining Grammar Formalism.
Anubaad	English to Bengali	CDAC, Kolkata	A hybrid system which uses n-gram approach for POS tagging. Works at sentence level
Anglabharti, AnglaHindi, Anubharti	English to Hindi, Tamil	IIT,Kanpur	Uses intermediate structure Pseudo Lingua for IL.
English Hindi MTS	English to Hindi	IIIT, Hyderabad	Combines Rule Based Machine Translation and phrase based SMT

Table 4: Machine translation systems in Indian languages

Speech to speech translation system has also been tried upon. CDAC Kolkata developed a prototype for Hindi-Bangla speech to speech dialogue system. A consortium project for speech to speech translation system was initiated at the international level and India was also a part of it (Arora et al.,2013).

4. Challenges in Technology Development for IL

In the race of language technology, Indian languages lie far behind the languages of other developed countries. The major requirement is for resource creation based on global standards. Apart from this, several other issues influence the technology growth for languages especially that of under resource languages. Some of the challenges faced by Indian languages for technology development is as follows :

- Language ambiguity and complexity: the same word has a different meaning when used in a different context
- Origin of Indian script and family: One language is represented using many scripts and also many languages follow the same script.
- Difficulty in data collection due to geographical, social and cultural strata of the country.

- Presence of several dialects: code-mixing between dialects; a massive number of non-native speakers of languages.
- Non-conformance with English centric models: existing models can't be extended to Indian languages.
- Localization issues associated with the operating system, keyboards and applications.
- Lack of encoding standards: several phones of Indian languages have not yet been encoded in existing standards

5. Way Forward for Indian Language Technology Development

Rigorous efforts are required to bring Indian languages on the world map of technologically developed languages. Some of the tasks are identified below to help curb the challenges faced in technology development for these languages:

- Producing a White paper: Necessary to reflect the current situation for all languages.
- Massive amount of text data creation to reliably train a statistical language model: focus should be on phonetically balanced data.
- Obtain transcribed recordings from several speakers to capture varying acoustical characteristics due to nativity and other sociolinguistics aspects for creation of the acoustic model.
- Pronunciation dictionary of the vocabulary for lexical/PLS development: the focus should be on capturing prosody.
- The urgency to work with zero resource language: use an AI approach and try to avoid its extinction.
- Generate facilities such as BLARK (basic language resource tool) for all IL.

6. Conclusion

Language technology contributes to promoting linguistic diversity and multilingualism in the digital world. Now, the technology is moving into the daily life of people in the different application area. India is a country with diverse linguistic variations. Very few Indian languages have been worked upon for the development of language technology. The present paper highlights the technological achievements of Indian languages. Many languages have shown their presence in the digital world and efforts in this direction is still continued. But, to date, the indigenous people are still experiencing barriers to access information through the internet. They experience obstacle to use a tool not available in native languages. Indian languages being under-resourced face more difficulty in this regard and may require a long term effort to get benefit from the latest digital developments.

7. Bibliographical References

Arora S., K. K. Arora, M. K. Roy, S. S. Agrawal, and B. Murthy, (2016). Collaborative speech data acquisition for under resourced languages through crowdsourcing," Procedia Computer Science, vol. 81, pp.37-44.

Arora, K., Arora, S. & Roy, M.K. (2013) Speech to speech translation: a communication boon. *CSIT* 1, 207–213 (2013) doi:10.1007/s40012-013-0014-4

Jha, G. N. (2010). India's language diversity and resources of the future: Challenges and opportunities. *Special Center for Sanskrit Studies, Jawaharlal Nehru University, New Delhi*.

Samudravijaya, K.(2000). Computer Recognition of Spoken Hindi. Proceeding of International Conference of Speech, Music and Allied Signal Processing, Triruvananthapuram, pages 8-13, 2000.

Singh, A., Kadyan, V., Kumar, M. *et al.(2019).* ASRoIL: a comprehensive survey for automatic speech recognition of Indian languages. *Artif Intell Rev* (2019) doi:10.1007/s10462-019-09775-8

S Sinha, S Sharan, S S Agrawal,(2017). O-MARC: A multilingual online speech data acquisition for Indian languages, Oriental-COCOSDA , Nov 1-3, 2017, held at Seoul, S Korea.

Sinha Shweta, Shyam S Agrawal (2018). Sustaining Linguistic Diversity Through Human Language Technology : A Case. Study for Hindi.May 2018. CCRUL-LREC 2018

Government of India, <www.censusindia.gov.in/ 2011Census/C-16_25062018_NEW.pdf> Accessed 08/01/2020

Governemnt of India, <tdil-.in/index.php?option= com_vertical & parentid =85 &lang=en> Accessed : 12/12/2019

Unesco, 2009 <www.unesco.org/languages-atlas/ index.php> Accessed 24/12/19