Development of Linguistic Resources and Tools for providing multilingual Solutions in Indian Languages – A Report on National Initiative

Swaran Lata , Somnath Chandra
Department of Information Technology
Ministry of Communications & Information Technology, Govt. of India
6 CGO Complex, Lodhi Road, New Delhi 110003
E-mail: slata@mit.gov.in, schandra@mit.gov.in

Abstract

The multilingual diversity of India is one of the most unique in world. Currently there are 22 constitutionally recognized languages with 12 scripts. Apart from these, there are at least 35 different languages and 2000 dialects in 4 major language families. It is thus evident that, development and proliferation of software solutions in the Indic multilingual environment requires continuous and sustained effort to edge out challenges in all core areas namely storage and encoding, input mechanism, browser support and data exchange. Linguistic Resources and Tools are the key building blocks to develop multilingual solutions. In this paper, we shall present an overview of the major national initiative in India for the development and standardization of Linguistic Resources and Tools for developing and deployment of multilingual ICT solutions in India. The initiative of Language Technology Development in India broadly categorized in three phases namely (i) Early Initiative-Seeding Effort, (ii) Consolidation Phase and collaborative development phase towards development of futuristic technology and associated phases. The National Initiatives in each phase of development and major achievement especially with respect to standardization of language resources and tools towards development of integrated language technology solutions are also highlighted in this paper.

1. Introduction

The world is in the midst of a technological revolution nucleated around Information and Communication Technology (ICT). Advances in Human Language Technology will offer nearly universal access to information and services for more and more people in their own language. India is multilingual country with 22 Constitutionally Recognized languages and 12 scripts. It is therefore essential that tools for information processing in local languages are developed and be made available for wider proliferation of ICT to benefit the people at large and thus paving the way towards "Digital Unite and Knowledge for all" and arrest the sprawling Digital Divide. The first step in this direction was initiation of Technology Development for Indian Languages (TDIL) Programme in year 1991, which has made available technology, tools and expertise in the country in the area of language technology.

Language technology development in India has now reached a stage today, where it has a potential to generate utility applications, benefiting the masses, which will enable people to access and use IT solutions in their common language.

It is well known that , Linguistic Resources and Tools are the key building blocks for development of integrated end-to-end multilingual solutions . Realizing the need for development of linguistic resources and tools , major national level initiatives have been undertaken for development and standardization of linguistic resources and tools in both written and speech domains.

Department of Information Technology (DIT) has further encouraged users and developers of Language Technology solutions by providing certain basic information processing tools like fonts, open office, e-mail client, internet browser, dictionary, conversion utilities, etc. free of cost, which will motivate users to use them to solve their basic problems and help developers to build advanced solutions. This will definitely boost up and leapfrog Indian language technology development and their deployment in a very fast way. In this paper we shall present a comprehensive overview of the national initiative for development of Linguistic resources and tools in Indian Languages. The paper is organized as follows. Section II would present a brief overview of the early initiatives. The Section III would describe the initiative of National Roll-Out Plan. Section IV describes the six of consortia projects in which linguistic resources and tools are being developed for providing Cross-lingual Information Retrieval, Machine Translation OCR, OHWR and Speech solutions in Indian languages. The Standardization Efforts for Indian Languages are described in Section V. The future directions and plans are presented in Section VI.

2. Early Initiative – The Seeding Effort

The first major development of Linguistic Resources and Tools have been nucleated in the year 2000 by establishing Resource Centres for Indian Language Technology Solutions (RC-ILTS) at 13 premier institutions for a period of three years in the year 2000. The idea was to proliferate this activity to a large number of institutions across the country with the specific mandate for a language or a group of languages. Under RC-ILTS initiative, several important linguistic resources and Tools have been developed. The major Linguistic Resources developed are:

2.1 Written Text Resources

(1) **Parallel Corpora:** One Million pages Parallel Corpora with graphical user interface in 13

languages namely Englsih, Hindi, Punjabi, Tamil, Telugu, Kannada, Malayalam, Bengali, Oriya, Marathi, Assamese, Gujrati and Nepali languages

(2) **Bi-lingual Dictionaries** Bi-lingual Dictionaries of English-Hindi, English-Bengali, English-Telugu, English-Tamil, English-Kannada, English-Malayalam, English-Oriya and Urdu-Hindi, each with over 30,000 root words

(3) **Ontology & Word-Net:** Hindi Word-net with 30000 sync-sets with morphological analyzer and front –end. Oriya Word-net with 1100 lexical entries with X-window interface.

(4) **On-line Vishwakosh** (Encyclopedia in Hindi) with 9162 topics

(5) **Phrasal Dictionaries** in Tamil and Kannada languages

(6) **Information Technology Terminology (10000 terms)** in Hindi

(7) **Text corpora** of 3 Million words in major Indian languages.

2.2 Speech Resources:

(1) **Speech Corpora:** Annotated Speech Corpora of approximately 50 hours has been developed for 10 Indian Languages namely Hindi, Marathi, Punjabi, Bengali, Assamese, Manipuri. Tamil, Malayalam, Telugu and Kannada languages The Speech Corpora with sample sound are available at Indian Language Data Centre http://www.lidc.gov.in.

(2) **Semi-Automatic Annotation Tool for Speech Corpora:** Semi-automatic annotation tool for speech database has also been developed Five levels of annotation namely phoneme, syllable, word, phrase and parts of speech (POS) are used for annotation. The Annotated speech signal and its output i.e. standard format time table (SFT) are available.

3. National Roll-Out Plan

The Linguistic Resources and tools developed during the seeding phase of the Language Technology development has been further consolidated, under the National Roll-Out Plan. Under the National Roll-Out Plan, CDs containing Software tools and fonts for all 22 Officially Recognized languages namely have been released in public domain for wider proliferation of benefits of Language Technology to masses.

All these software tools and fonts can be downloaded free from Indian Language Data Centre portal http://www.ildc.gov.in. So far approximately 4.0 million downloads have been recorded from this web site. CDs are also shipped free of cost to the users on request. So far approximately 500,000 CDs are shipped.

The basic contents of each of the Language CDs are elucidated in table below:

True Type Fonts with Keyboard Driver - more than 200	Supporting INSCRIPT, Typewriter, Phonetic Keyboard layouts Allows content creation in Indian languages using applications running under Microsoft windows
Language Multi-font Keyboard Engine for True Type Fonts	Allows content creation in Indian languages using applications running under Microsoft windows in variety of font encoding.
Language Unicode Compliant Open Type Fonts - more than 200	Allows rendering the Indian language Unicode data.
Unicode Compliant Keyboard Driver	Supporting INSCRIPT, Typewriter, Phonetic Keyboard layouts. Allows Unicode complaint data inputting
Generic fonts and storage code converter	Allows user to convert the existing data in different encoding to ISCII / UNICODE
Localized version of Open Office	This consists of word processor, presentation tool, spreadsheet & drawing tool
Fire fox browser	Localized version of Fire fox browser
GAIM	Multi-protocol Messenger. This enables the user to user various messenger clients for communications
Typing Tutor	This application teaches the user to type in Indian languages
Spellchecker	Allows the end user to rectify spelling mistakes in the document for Indian Languages
Dictionaries	English to Indian language and vice versa dictionaries in general, administrative, technical domains.
Transliteration Tool	Transliterates a given Indian language text into Roman & vice versa. Useful for user who is not familiar with the script.

Table 1: Contents of CD

Some of Indian Languages Tools and Applications are depicted in Figures 1, 2, 3 respectively.

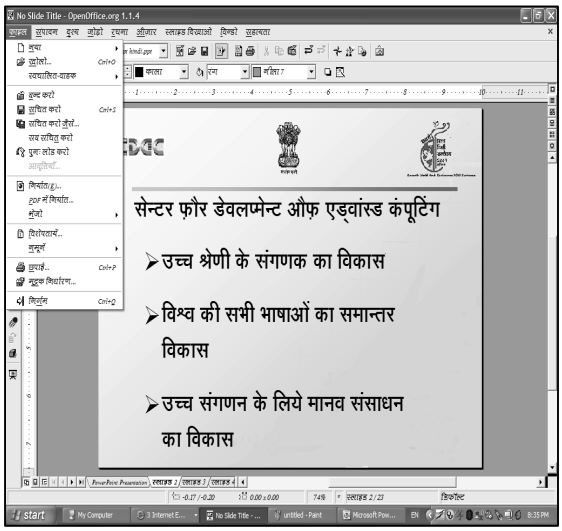

Figure 1: Hindi Open Office

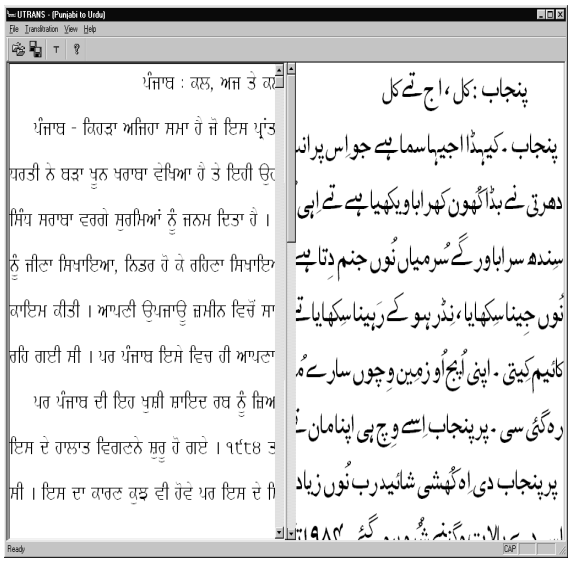

Figure 2: Punjabi-Hindi Transliteration Tool

Figure 3: Sample Hindi Application

4. Consortium Mode Projects

The earlier efforts in the seeding phase as mentioned earlier resulted in the development of Linguistic resources and tools in a scattered and non-standard manner. In order to develop Standardized language technology solutions for mass impact applications, consortium mode projects have been initiated in the areas of English to Indian Languages Machine Translation , Indian languages to Indian languages Machine Translation , Cross-lingual Information Access, (CLIA) , Optical Character Recognition , and On-line handwriting Recognition systems , Text to Speech in Indian Languages and Automatic Speech Recognition. Seven Consortia involving Thirty seven academic and research institutions are working in these areas. In addition to the above mentioned consortia, three other consortia are working specifically in the areas of Indian Language Annotated Corpora, Word-Net and Speech Corpora. Several Linguistic Resources and Language Tools are being developed in these consortia. Some of the notable resources are:

1) Multilingual Sense Dictionary in 6 Indian Languages Pairs

2) Morphological Analyzers for major Indian Languages

3) Font Trans-coder

4) Indian Languages Annotated Corpus for Tourism Domains for 11 Indian Languages

5) Word-net for Indian Languages Hindi , Marathi, Assamese, Bodo , Manipuri and Nepali languages. It will be linked with English Wor-Net.

6) Speech Corpus in 6 Indian Languages Hindi, Bengali, Tamil, Telugu, Malayalam and Marathi languages for Text-to-speech system applications.

In addition to the above, new consortia mode projects are also being initiated for development of :

(1) Pronunciation Lexicon as per W3C specifications (PLS1in major Indian languages

(2) Duration Models for Indian Languages speech.

5. Standardization Efforts

Due to the complexity of the Indian Languages and scripts, it is also essential to develop standards and best practices for Data storage, Input mechanism, Locale data and Data Exchange formats. Some the major initiatives are enlisted below:

5.1 UNICODE:

Department of Information Technology is the voting member of the Unicode Consortium to ensure the adequate representation of Indic scripts in the

Unicode Standards. DIT finalized the changes in the Unicode Standard and majority of changes have been accepted and incorporated in Unicode Standards version 5.1. Sixty five characters have been recommended for inclusion in the Unicode Standard for representation of Vedic Sanskrit in Unicode. Also, 56 characters as recommended by Government of Manipur have been recommended for encoding for representation of Manipuri language in *Meetei Mayek* script. Initiatives have also been undertaken to encode *Grantha* script.

A snapshot of the effort is depicted in figure 4 below:

Signs for Yajurvedic

1CD5	◌᳕	VEDIC TONE YAJURVEDIC AGGRAVATED INDEPENDENT SVARITA = vaidika svarita adho nyubja	1CEA
1CD6	◌᳖	VEDIC TONE YAJURVEDIC INDEPENDENT SVARITA = vaidika svarita addah konna	1CEB
1CD7	◌᳗	VEDIC TONE YAJURVEDIC KATHAKA INDEPENDENT SVARITA = vaidika svarita adho vakra rekha	1CE8
1CD8	◌᳘	VEDIC TONE CANDRA BELOW = vaidika svarita adhah ardha vakra	1CE9
1CD9	◌᳙	VEDIC TONE YAJURVEDIC KATHAKA INDEPENDENT SVARITA SCHROEDER *=vaidika svarita adho samyukta rekhaa*	1CEE
1CDA	◌᳚	VEDIC TONE DOUBLE SVARITA = vaidika svarita uurdhva dvi rekha	1CE5
1CDB	◌᳛	VEDIC TONE TRIPLE SVARITA = vaidika svarita uurdhva tri rekha	1CE6
1CDC	◌᳜	VEDIC TONE KATHAKA ANUDATTA = vaidika svarita adho rekhaa	1CEC
1CDD	◌᳝	VEDIC TONE DOT BELOW = vaidika svarita adho bindu	1CED

Figure 4: Unicode for Vedic Sanskrit

5.2 Common Locale Data Repository (CLDR): Modifications/ Development of UNICODE Common Locale data repository (CLDR) containing major fields dates, times, time zones, numbers, and currency values; sorting text; etc. in Indian languages have been initiated in consultations with state governments and other stake holders.

5.3 Language Tags: The present forms Language Tag of ISO 639-2, ISO 639-3 and the futuristic ISO 639-5 and ISO 639-6 have many ambiguous entries for Indian languages, which need to be corrected urgently in order to prevent propagation of incorrect nomenclatures for Language sets. Modification / Additions of Language Tags in Indian languages have been taken up in consultations with the state governments and all stake holders.

5.4 W3C Web Standards :
Major Initiative has also been undertaken for adequate representation of Indian Language Specificities in W3C existing and futuristic web standards. Initiatives have been taken for incorporation of Indian languages requirements in W3C Standards based on the national priority.

The verticals of work chosen are:

- Internationalization
- Web Architecture
- Mobile Web Initiative
- Styling including CSS
- Web Services ,Web Applications and Service Modeling Language
- XML , XML associated standards and Efficient Interchange of XML
- Speech and associated Mark-up Languages
- E-Government
- Semantic Web including OWL and RDF

Focus is being given on specific requirements of Indian Languages because of its non-linearity for incorporation in specific W3C Standards. Work has been initiated for development of Pronunciation Lexicon in Indian Languages.

5.5 Script Grammar:
The nonlinear nature of Indian Scripts requires standardization of Script Grammars. Initiatives have also been undertaken for standardization of Script Grammars for all 22 Official languages.

Initiatives have also been undertaken for standardization of Domain Names in Indian Languages and IPA representation for Indian Languages.

6 Future Directions

The complexity and vastness of Indian Language Ecosystem requires sustained and collaborative efforts for development and standardization of Linguistic Resources and Tools towards development of ICT solutions in Indian Languages. Comprehensive policy for standardization, testing and evaluation of Linguistic Resources and Tools are being planned. Testing and Evaluation campaigns inline of those of international efforts like NIST CLEF etc.

7 Conclusion

In this paper we have presented an overview of the development and standardization of Linguistic resources and tools for complex Indian Multilingual environment. We conclude with the comment that, the challenges for implementation of Multilingual ICT solutions in India are huge, but opportunities are unlimited.

References:

[1] http://www.ildc.gov.in
[2] http://tdil.mit.gov.in