Challenges of Multilingual Web in India: Technology Development & Standardization Perspective

Swaran Lata
Country Manager, W3C India & Head, TDIL Programme, Dept of Information Technology, Govt of India
E-mail: swaran@w3.org, slata@mit.gov.in
Organization of my talk:

- **India**
  - Internet Scenario in India
  - Multilingual Web based service in India
  - Major web based services in India
    - E-Governance – Common Service Centers & Citizen centric services
    - Unique ID Project – Aadhar

- **Multilingual Complexity In India:**
  - Origin of Indic Scripts & family
  - Complex Characteristics

- **Technology Challenges**
  - Ab-initio development of Language Computing Solution
  - Non-Conformance with English –Centric Model
  - Complex nature of key technologies (Machine Translation, Cross-lingual Information Access, Character Recognition etc)
  - Some key Initiatives

- **Standardization Challenges:**
  - Core:
    - Encoding
    - Input
    - Display
  - Middleware: HTML, CSS, Mobile Web, Web Accessibility & Mobile Web
  - Application Level: Presentation Issues & Web design implementation issues
Internet Usage Statistics in India
Wireless subscriber Statistics in India

- Total Telephone subscriber base reaches 706.37 million
- Wireless subscription reaches 670.60 million
- Wireline subscription reaches 35.77 million
- Overall Tele-density reaches 59.63
- Broadband subscription is 10.08 million

![Wireless Subscriber Base Chart](chart.png)
E-Governance – Common Service Centers & Citizen centric services

The Government of India has launched the National e-Governance Action Plan (NeGP) for G2G,G2B,G2E and G2C services.

Modalities for Government on the Web

- Provide: public services on the web, either transactional or information services or both.
- Engage: with citizens and businesses, on government terms or on the citizens terms.
- Enable: public sector information re-use.
Web based services in India

Panchayati Raj

Department of Agriculture & Cooperation

Bengal state Portal

Kerala IT Mission
Web based services in India

Gujarat State Portal

Tamil Nadu Government portal

Jan Mitra

E-Choupal
Online Services under National e-Governance Plan

Central Mission Mode Programme (MMP)s

- Income Tax
- Passport/VISA
- Company Affairs
- Central Excise
- Pensions
- Land Records
- Road Transport
- Property Registration
- Agriculture
- Municipalities
- Police
- Employment Exchange
- E-Courts

State Mission Mode Programme (MMP)s

- Agriculture
- Commercial Taxes
- e–District
- Employment Exchange
- Land Records
- Municipalities
- Gram Panchayats
- Police
- Road Transport
- Treasuries
Nation-wide Unique ID Project

- The Unique Identification Authority of India (UIDAI) was established in 2009 by the Government of India, with the developmental mandate of setting up the infrastructure to provide a universal way of uniquely identifying Indian residents.

- AADHAAR, a 12-digit unique identification number (UID) that will be provided after getting the demographic and biometric information of an individual.

- AADHAAR's guarantee of uniqueness and centralized online identity for building the multiple services and applications

- AADHAAR give ability to access these web based services and resources, anytime, anywhere in the country

Aadhar – Towards foundation of Nation-wide multilingual web based services accessible to its citizens irrespective devices & platforms
Host of On-line Nation-wide Web Based Services for a population over 1 billion

Multilingual Data base Development

Multilingual On-line Services

Web Interface for all including elderly and physically challenged sections
Multilingual Complexity In India
Languages of India

- According to Census 2001 India has 122 major languages and 2371 dialects.
- Out of 122 languages 22 are constitutionally recognized languages.
- Linguistic Diversity is very rich and wide in India
- One Language –many scripts
- Many Language –one script
- Culturally different depending on region though using same script for different languages.
- Even wide difference for same language across different parts of the country
<table>
<thead>
<tr>
<th>States</th>
<th>Languages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chhattisgarh</td>
<td>Hindi</td>
</tr>
<tr>
<td>Himachal Pradesh</td>
<td>Hindi</td>
</tr>
<tr>
<td>Arunachal Pradesh</td>
<td>Assamese</td>
</tr>
<tr>
<td>Karnataka</td>
<td>Kannada</td>
</tr>
<tr>
<td>Kerala</td>
<td>Malayalam</td>
</tr>
<tr>
<td>Madhya Pradesh</td>
<td>Hindi</td>
</tr>
<tr>
<td>Maharashtra</td>
<td>Marathi</td>
</tr>
<tr>
<td>Manipur</td>
<td>Manipuri (Meitei)</td>
</tr>
<tr>
<td>Mizoram</td>
<td>Mizo, English</td>
</tr>
<tr>
<td>Nagaland</td>
<td>English</td>
</tr>
<tr>
<td>Orissa</td>
<td>Oriya</td>
</tr>
<tr>
<td>Punjab</td>
<td>Punjabi</td>
</tr>
<tr>
<td>Rajasthan</td>
<td>Sikkim</td>
</tr>
<tr>
<td>Tamil Nadu</td>
<td>Tamil</td>
</tr>
<tr>
<td>Tripura</td>
<td>Tripura</td>
</tr>
<tr>
<td>Andhra Pradesh</td>
<td>English</td>
</tr>
<tr>
<td>Chandigarh</td>
<td>Hindi</td>
</tr>
<tr>
<td>Goa</td>
<td>Konkani</td>
</tr>
<tr>
<td>Maharashtra</td>
<td>Gujarati</td>
</tr>
<tr>
<td>Haryana</td>
<td>Hindi</td>
</tr>
<tr>
<td>Jharkhand</td>
<td>Hindi</td>
</tr>
<tr>
<td>Lakshadweep</td>
<td>English</td>
</tr>
<tr>
<td>Meghalaya</td>
<td>Khasi, Garo</td>
</tr>
<tr>
<td>Uttar Pradesh</td>
<td>Assam</td>
</tr>
<tr>
<td>West Bengal</td>
<td>Assam</td>
</tr>
<tr>
<td>Assam</td>
<td>Assamese</td>
</tr>
<tr>
<td>Bihar</td>
<td>Maithili</td>
</tr>
<tr>
<td>Damra and Nagar Haveli</td>
<td>Gujarati</td>
</tr>
<tr>
<td>Daman and Diu</td>
<td>Gujarati</td>
</tr>
<tr>
<td>Delhi</td>
<td>Bodo</td>
</tr>
<tr>
<td>Jammu and Kashmir</td>
<td>Urdu</td>
</tr>
<tr>
<td>Puducherry</td>
<td>Tamil</td>
</tr>
<tr>
<td>Uttarakhand</td>
<td>Hindi</td>
</tr>
<tr>
<td>Andaman and Nicobar Islands</td>
<td>Hindi</td>
</tr>
<tr>
<td>Andhra Pradesh</td>
<td>Telugu</td>
</tr>
<tr>
<td>Telugu</td>
<td>Hindi</td>
</tr>
<tr>
<td>Telugu</td>
<td>Hindi</td>
</tr>
<tr>
<td>Tamil</td>
<td>Tamil, Telugu</td>
</tr>
<tr>
<td>Punjabi</td>
<td>Dogri</td>
</tr>
<tr>
<td>Utah</td>
<td>Dogri</td>
</tr>
<tr>
<td>Telugu</td>
<td>Telugu</td>
</tr>
<tr>
<td>Bengali</td>
<td>Urdu</td>
</tr>
<tr>
<td>Hindi</td>
<td>Hindi</td>
</tr>
<tr>
<td>Urdu</td>
<td>Hindi</td>
</tr>
<tr>
<td>Dogri</td>
<td>Urdu</td>
</tr>
<tr>
<td>Malayalam</td>
<td>Sanskrit</td>
</tr>
<tr>
<td>Sanskrit</td>
<td>Bengali</td>
</tr>
</tbody>
</table>
**Major Scripts and Corresponding Languages in India**

**Northern Scripts** (Gupta Scripts)
- Brahmi Script (Ashokan)
- Indus Script (proto Brahmi Scripts)
- Unknown Ancient Scripts

**2000 BC**

**400 BC**

**3rd BC**

**20th Century**

- Landa
- Sharda
- Kutil

**7th century**

- Gurmukhi
- Sardar

**8th Century**

- Gaur
- Nagari
- Kauri
- Kaithi

**10th Century**

- Devanagari
- Gujarati

**Unknown Ancient Scripts**

**Southern Scripts**

- Brahmi Script (Ashokan)
- Kharoshthi Script 400 BC-300 BC
- 3rd Century
- 7th Century
- 14th Century

- Tamil
- Malayalam
- Southern Sinhalese
- Central Sinhalese
- Grantha

- South-eastern Asian-Burmese, Thai, Cambodian, Indonesian, Malasiyan, vietnam, Philipines etc
- Telugu
- Kannada
- Vettashut
- Kolehat

- Assamese
- Maithali
- Central Asian
- Ol-Chiki
- Nepali (Newari)

- Bangla
- Maithali
- Tibetan

- Ol-Chiki
- Nepali (Newari)

- Maithali
- Ol-Chiki
Technology Challenges
Technologies Developed under consortium mode projects

• English to Indian Languages Machine Translation System
  [6 Language Pairs: English to Hindi, Marathi, Bengali, Oriya, Tamil, Urdu.]

• Indian Languages to Indian Languages Machine Translation System

• Cross-Lingual Information Access
  [6 Languages : Hindi, Bengali, Tamil, Marathi, Telugu and Punjabi]

• Optical Character Recognition Systems
  [10 Scripts: Bangla, Devnagari, Malayalam, Gujrati, Tamil, Telugu, Kannada, Oriya, Gurumukhi, Tibetan]

• On-line Handwriting recognition system
  [6 Scripts: Hindi, Bengali, Tamil, Telugu, Kannada and Malayalam]
Sample Outputs For English – Hindi

Complex nature of key technologies: Machine Translation

English To Indian Languages Machine Translation System
Funded by: TDIL Program, Department of Information Technology, Govt. of India

You are: user42 Language: Hindi

* This page can display maximum of 5 outputs at a time.

<table>
<thead>
<tr>
<th>Input</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>The best time to visit Jaipur is December to February</td>
<td>जयपुर दर्शन करने के लिए सर्वोत्तम समय दिसंबर से फरवरी है</td>
</tr>
<tr>
<td>Bara Shingri Glaciers is the largest in Himachal Pradesh</td>
<td>बारा शिंग्री ग्लेशियर हिमाचल प्रदेश में सबसे बड़ा है</td>
</tr>
<tr>
<td>The Jal Mahal is a picturesque palace built for royal duck shooting parties</td>
<td>जल महल शाही बतख आकर्ष़ प्रदेश के लिए निर्माण किया गया जीवन्त महल है</td>
</tr>
</tbody>
</table>

Developed by Consortium of Institutions - CDAC Pune, IIT Bombay, CDAC Mumbai, Jadavpur University, IISc-Banglore, Utkal University, Banasthali Vidyapeeth, Amrita University, IIIT Allahabad, IIIT Hyderabad
### English To Indian Language Translation System

**User:** Mohan  |  **Language:** Urdu

---

**PrintTargetText | PrintEnglishText | RebuildTargetText | Translation | SignOut

---

**Final Output**

**English:**

*The Amber fort complex has several apartments with palaces, halls, stairways, pillared pavilions, gardens & temples.*

*Amber palace is a classic example of Mughal & Hindu architecture.*

**Urdu:**

*امبر قلعہ کے مختلف محلوں، دیواریں، سیکڑوں سے تکریم، باغات، ستون، سرسبز گنبد، پیلار، پروں، چھوٹے گھر، جنگلی چھوٹے گھر، مغل و هندی بنیاد کے نمونے*
The Amber fort complex has several apartments with palaces, halls, stairways, pillared pavilions, gardens & temples.

Jaipur, popularly known as the Pink City, is the capital of Rajasthan state, India.

আম্বার পুরো সম্প্রসারিত প্রাসাদ, বড় প্রাসা, সিড়ি, কুঠিয়া প্রাসাদ, বাগান এবং অনিস্ত বিভিন্ন কৃষ্ণকুণ্ড কাশ্মীর উৎসাহ রাজস্থানের রাজধানী।
Sample Outputs For English - Tamil

April to June and October to November are the best time to visit the town.

Guru Shikhar is the highest peak on the Mount Abu which provides an excellent view of the whole town.
Indian Language to Indian Languages Machine Translation System

Available experimental language pairs through Sampark
- Punjabi to Hindi (version-1.1.0)
- Hindi to Punjabi (version-1.0.0)
- Telugu to Tamil (version-1.0.2)
- Urdu to Hindi (version-1.0.7)

After login click Duur Sampark for these languages
- Hindi to Telugu (version-1.0.8)
- Tamil to Hindi (version-1.0.6)
- Hindi to Urdu (version-1.0.1)

* Consortium of Institutions - IIIT Hyderabad, University of Hyderabad, CDAC (Noida, Pune), Anna University KBC Chennai, IIIT Kharagpur, IIT Bombay, IISC-Banglore, Tamil University, IIIT Allahabad, Jadavpur University
Urdu-Hindi ILMT Translation through URL:

(Source web page: http://www.bbc.co.uk/urdu/entertainment/2010/02/100212_my_name_khan_ka.shtml)
Cross Lingual information access integrated with Machine Translation

English Database Crawling

English Query

Input processing [Query translation /Transliteration]

Searching & Indexing

English Output search result

Machine Translation Service

Eng – Indian Languages

Ind. Lang Output search

Query in Indian Languages
In CLIA, the input query is in one language and information is retrieved in another.

The query language is one of Bangla, Hindi, Marathi, Punjabi, Tamil and Telugu.

The retrieved documents are in English, Hindi or the language of the query.
বঙ্গা মনোলিঙ্গুয় রিট্রিভেশন

Bengali Monolingual Retrieval
Bengali – Hindi Cross-lingual Retrieval
अब अदालत के फैसला सुनने की बारी आयी। सभी को रमा से सहानुभूति  
हो पायी थी, पर इसके माध्यम से यह भी मानी हुई बात थी कि उसे सत्य होगी।  
क्या सत्य होगी, वही देखना था। लोग बही उस्तुला से फैसला सुनने के लिए  
और सिमट आये, कुस्तियां और एक्साय ली गयीं, और कन्नवातिया भी बद रहीं  
गयी।  
मुख्यों के समन कहते हैं कि एक चुंबक ने अपनी प्राण-रक्षा के लिए पुलिस  
का आतंक लिया और उसे लाल बनाया था कि तक पहुंचे वह पुलिस का  
अटक नहीं रहे। जब वह लाल बना है तो यह बात हो जाती है कि तक पहुंचे  
वह ऐसा रहता है, अर यह रेखाओं जो जांच के लिए लाल के लिए लाल  
पुकार होता है। उन्होंने आगे चलकर झुंड के दो हाथ ने पुकार का  
तंबाकू चढ़ाया। हां, वह लाल बना है तो यह बात हो जाती है कि  
वह ऐसा रहता है, जब वह लाल बने होते हैं। उन्होंने आगे चलकर झुंड  
के दो हाथ ने पुकार का तंबाकू चढ़ाया। हां, वह लाल बना है तो यह  
बात हो जाती है कि वह ऐसा रहता है, जब वह लाल बने होते हैं। उन्होंने  
आगे चलकर झुंड के दो हाथ ने पुकार का तंबाकू चढ़ाया। हां, वह लाल  
बना है तो यह बात हो जाती है कि वह ऐसा रहता है, जब वह लाल  
बने होते हैं। उन्होंने आगे चलकर झुंड के दो हाथ ने पुकार का  
तंबाकू चढ़ाया। हां, वह लाल बना है तो यह बात हो जाती है कि  
वह ऐसा रहता है, जब वह लाल बने होते हैं। उन्होंने आगे चलकर झुंड  
के दो हाथ ने पुकार का तंबाकू चढ़ाया। हां, वह लाल बना है तो यह  
बात हो जाती है कि वह ऐसा रहता है, जब वह लाल बने होते हैं। उन्होंने  
आगे चलकर झुंड के दो हाथ ने पुकार का तंबाकू चढ़ाया। हां, वह लाल  
बना है तो यह बात हो जाती है कि वह ऐसा रहता है, जब वह लाल  
बने होते हैं। उन्होंने आगे चलकर झुंड के दो हाथ ने पुकार का  
तंबाकू चढ़ाया। हां, वह लाल बना है तो यह बात हो जाती है कि  
वह ऐसा रहता है, जब वह लाल बने होते हैं। उन्होंने आगे चलकर झुंड  
के दो हाथ ने पुकार का तंबाकू चढ़ाया। हां, वह लाल बना है तो यह  
बात हो जाती है कि वह ऐसा रहता है, जब वह लाल बने होते हैं।
திருக்குருக்கள் திருமால் பாடல்

பாடலால் மாறியுள்ள உண்மையான ராகத்து வைத்தே பலரால் மூண்டப்படுத்துக! அன்னு அம்மனை திருப்போக்கு பலரால் பிளிய மருத்துவ யாவது விளக்கம்? பச்சிலடைந்து பார்த்தேன்று விளக்கத்து மூன்று பிளிய மருத்துவ யாவது? திரும்பு புண்டம்பான அப்பாச்சுக்கு வாழ்க்கை திறன்களே, திறன்களே, அந்தார், பலராகவே, தமிழிலிருந்து வாழ்க்கை, தமிழிலிருந்து விளக்கத்து, அன்னு புண்டம்பான.
Output of Tamil OCR
Sample Tamil OHWR form - Tamil
Speech Technology:

Speech Corpus:
Annotated Speech Corpora of approximately 50 hours developed for Hindi, Marathi, Punjabi, Bengali, Assamese, Manipuri, Tamil, Malayalam, Telugu, Kannada.

Text-to-Speech for 6 Indian Languages Hindi, Bengali, Marathi, Tamil, Telugu and Malayalam is being developed for visually challenged section of society.

Speech Recognition:
Consortium Mode Project has been initiated for development of Automatic Speech Recognition systems for agricultural commodity prices in six Indian Languages: Hindi, Tamil, Telugu, Bengali, Assamese and Marathi languages.

Phonetic Engine for Speech recognition system for Hindi and Telugu languages are being developed.
Sample Waveform for Tamil Speech: Phonetic Engine

/talaivarAha tErndeDukkappattaipin pradinidihaLiDam/
(கற்று புக்கா மைக்கப்பட்டின் பரிதேய்க்கில்)
Standardization Aspects for Supporting Multilingual Web in Indian Languages
Character Encoding: UNICODE

- Basis of Multilingual Web.
- All data exchange would be possible seamlessly across devices and Platforms.
- Unicode Encoding for all 22 Constitutionally Recognized Indian Languages Complete.
- **Unicode declared as a Standard for Data Storage for Web Based E-Governance Services in India**
Styling Issues in Indic Languages
Drop Letters in Indian languages

- Issues for Indian Languages with respect to first character used in Hindi, Malayalam, Bengali, Telugu and Gujarati etc.
Underlining of characters

- There is some examples of Indian languages in which Matra’s are not readable due to underlining of characters

Hindi - अन्य भाषाओं में भी अनुवाद

Punjabi - ਬਾਂਧ

Bengali - তাই পুরানো আর্কাইভ একটি ওলট পালট।

Gujarati - सरदार गुजरती

Marathi - मराठी मुला मुल्लिची नावे

Tamil - நின்றுவாடியும் விளக்கம் திகழ்த்துவடா

Telugu - తెలుగు టెలివిషన్ టెల్ 9 తెల్లు ఇదో తెలు ఇదో లోతు
- **Vertical arrangements**

- **Bullets and Numbering**

- **Indentation of character**
Under Lining of the characters

Hindi: <p>राष्ट्रमंडल खेलों के लिए स्टेडियम सीजन को तैयार नहीं है.......</p>

English: <p>Commonwealth Games stadium is not ready for handover......</p>

Bengali: <p>পর্যটন অন্তর্ভুক্ত করা হয়েছে।</p>

Tamil: <p>பாராட்டி அனுப்பது, பொறியல்: அட்வார் பாதுகாப்பு......</p>

Gujarati: <p>સ્ટેડર્ટની સામાજિક પ્રબંધનથી, પ્રશાસકીય પ્રતિષ્ઠાની સામાન્ય સુરક્ષા......</p>

Marathi: <p>बंदोबस्त धरण तेजबुद्ध ते समधवर्ती चाँद वापसी गांठ राजकीय कार्य......</p>

Punjabi: <p>ਸਾਲ ਦੀਆਂ ਕੰਠੇ ਦਿੱਲੀ ਲਈ ਰਾਜਧਾਨੀ ਰਾਖ ਲਗਾਂ ਰਾਜਕੀ ਕੰਟਰੋਲ.....</p>

Malayalam: <p>സഞ്ചാര പ്രവൃത്തി പരിശീലനം നടത്താൻ തയ്യാരായ ഉപകരണങ്ങൾ......</p>

Kannada: <p>ಹಿಂದಿ ಸೈಬರ್ ಸೇವೆಗಳಿಗೆ ಬಳಸಲಾಗಿದೆ.....</p>

</body>
</html>
Major Identified Problems in Styling:

- Grapheme Cluster Problems for Vertical Writing Style
- Drop Initial Views of the First Letter Element
- Bullets & Numbering issues
- Justification Problems
- Horizontal Letter Spacing Problems

Most browsers are unaware of syllable boundaries for Indic scripts.

Current workaround for this is enclosing the first syllable in a separate span governed by a dedicated CSS class.

Mozilla Firefox seems to recognize it well.
Approach to be taken for Possible Solution

1. Grapheme Cluster Problems:
   Adoption of UAX#29 (Unicode Text Segmentation Algorithm addressing the complexities of Indian Grapheme Clusters)

2. Bullets & Numbering issues

3. Justification Problems

4. Horizontal Letter Spacing Problems

   - Development of Complete Mapping Table involving detailed requirements for document layout, typography and typesetting and calligraphic conventions, i.e. Styling Manual.
   - The Rules developed in Styling Manual needs to be converted for inclusion in HTML and CSS
• An activity involving in-depth analysis of browser rendering engine needs to be taken up on a urgent basis.

• A reference web-page layout engine needs to be developed which could act as a benchmark for other browser players to emulate.

• Mozilla Firefox could be taken up for modification to act as a reference implementation.

• Core Working Group under the aegis of W3C Internationalization Group being formed.
Enabling speech interface for web
Indian language requirement in Speech Interface

Diagram:
- User
- PLS Document
- Speech Recognition (ASR)
- SRGS
- Speech Synthesis (TTS)
- SSML
- Language Interpretation
- Language Generation

Requirements in Indian languages
W3C Pronunciation Lexicon Specification (PLS 1.0) – Voice Browsing

W3C Voice Browser Activity, has published a Pronunciation Lexicon Specification (PLS) Version 1.0. PLS is designed to enable interoperable specification of pronunciation information for both speech recognition and speech synthesis engines within voice browsing applications.

**PLS Attributes:**

**Multiple pronunciations**
- For ASR: to accommodate speaker/regional variability, not native speakers
- For TTS: a single preferred pronunciation will be selected

- **Multiple orthographies (with same pronunciations)**
- Useful for both ASR & TTS

- **Homophones (same pronunciations, different meanings)**

- **Homographs (same spellings, different pronunciations)**
Proposed Approach for incorporating Multilingual requirement in PLS:

**Usage of Parts of Speech (POS) information for resolving multiple pronunciations**

- Parts of Speech (POS) plays an important role in Indian languages (like Bangla, Hindi) pronunciation.

- Based on the POS same orthography can produce different pronunciation.

**Pos as an attribute:**

```xml
<?xml version="1.0" encoding="UTF-8"?>
<lexicon version="1.0" xmlns="http://www.w3.org/2005/01/pronunciation-lexicon"
    alphabet="ipa" xml:lang="bn">
    <lexeme>
        <grapheme>সরল</grapheme>
        <phoneme pos="adjective">fɔɾəl</phoneme>
        <!-- IPA string is: "fɔɾəl" -->
        <!--Itrans is: "sarala" -->
        <!--Meaning is : "easy" -- >
        <phoneme pos="verb">ʃɔɾlo</phoneme>
        <!-- IPA string is: "ʃɔɾlo" -->
        <!--Itrans is: "sarala" -->
        <!--Meaning is: "moved" -- >
        <phoneme pos="null">fɔɾəl</phoneme>
        <!-- IPA string is: "fɔɾəl" -->
        <!--Itrans is: "sarala" -->
        <!--Meaning is: "easy" -- >
    </lexeme>
</lexicon>
```
Usage of morphological information for resolving multiple pronunciations

- In Indian Languages not only POS information but also morphological information are very crucial in determining the pronunciation of a homograph.

- This information can be defined in the same attribute or element used for POS using proper POS tag set for that language.

Example: Bengali

```xml
<?xml version="1.0" encoding="UTF-8"?>
<lexicon version="1.0" xmlns="http://www.w3.org/2005/01/pronunciation-lexicon"
     alphabet="ipa" xml:lang="bn">
    <lexeme>
        <grapheme>করে</grapheme>
        <phoneme: pos="VM.3.prs.smp.dcl.fin.n.n.n">করে</phoneme>
        <!-- IPA string is: "kore" -->
        <!-- Trans is: “kare” -->
        <!-- Meaning is: “do/does” -->
        <phoneme: pos="VM.0.0.0.0.mfn.n.n.n">করে</phoneme>
        <!-- IPA string is: "kore" -->
        <!--  Trans is: “kare” -->
        <!-- Meaning is: “having done” -->
    </lexeme>
</lexicon>
```
Web Accessibility & Implementation
### Challenges:

- Make Web content accessible to people with disabilities w.r.t Indian languages
- WCAG 2.0 Guidelines for success criteria vis-a-vis selected recommendations relevant to Indian context

### Initiatives in India:

- “Guidelines for Indian Government websites” by NIC, Govt. of India
- STQC Implementing WCAG 2.0 Accessibility through Website Quality Certification
- Centre for Internet and Society developing authorized translation of WCAG 2.0 Guidelines

### Roadmap:

- Meet WCAG 2.0 guidelines & techniques w.r.t Indian languages
- Initiation with Hindi, Bangla, Marathi, Telugu, Tamil
### Some of the WCAG 2.0 Compliant Web Sites

<table>
<thead>
<tr>
<th>Name of the Portal</th>
<th>Snap Shot</th>
<th>Level of Compliance</th>
</tr>
</thead>
</table>
| [www.socialjustice.nic.in](http://www.socialjustice.nic.in)  
(Ministry of Social Justice & Empowerment) | ![Snapshot](image1.png) | AA                  |
| [www.mit.gov.in](http://www.mit.gov.in)  
(Depart of Information Technology)       | ![Snapshot](image2.png) | AA                  |
Issues for enabling Mobile Web in Indian languages

- Character encoding
- Bandwidth and Cost
- Backward Compatibility with Legacy Devices
- Lack of standardization
- Fonts
  - Bit map fonts (used by low cost handset)
  - True type fonts (used by high end handsets)
  - Open type fonts (currently in wider use)
- Common Storage format
- Mobile messaging in indic languages
- Presentation Issues
Mobile keypads issues

- **Multi-tap issues**
  - Too many taps per key for each char No way to know which char is on which Key.
  - Never support more than one language on the keypad because there is not enough space on the key face to print more characters.

- **Dictionary Based**
- **Transliteration**
Level of Adherence to W3C Mobile Ok while testing 10 Hindi websites

- 7 websites with 0-20% W3C validation level
- 2 websites with 21-40% W3C validation level
- 1 website with 41-60% W3C validation level
Proposed approach: Gap Analysis

W3C MWI
- MWI Standardization activity
- W3C Mobile OK
- Mobile ok tests suites
- Default Delivery context for Mobile applications

Indian requirements
- MWI Standardization requirements in India
- Implementation of Mobile OK for Indian languages
- Implementation of Mobile ok tests suites for ILs
- Need to update Default Delivery context for Mobile applications in ILs
Development of Recommendations for Mobile Ok in Indian Languages

W3C Current

Scope  MWBP  Mobile OK Scheme  Mobile OK Checker  Mobile OK Tests

Requirements of Indian Scenario

GAP Analysis

Scope  MWBP  Mobile OK Scheme  Mobile OK Checker  Mobile OK Tests

Observation, Recommendation for Future activity
Future Initiatives
Semantic Web

1. Development of Word-Net for Indian Languages.

2. Development of Multimedia Web Ontology Language (MWOL)

- **Multimedia Web Ontology Language** (MOWL), is designed as an extension of OWL, the W3C recommended ontology language for the web.

- **MOWL** supports creation of and reasoning with perceptual modeling of concepts, and probabilistic evidential reasoning.

Formation of the concept “Medieval Indian Monument” and the abstracted visual patterns that are expected on an embodiment of the concept in a multimedia artifact.
Some of Future Initiatives:

- Multilingual Requirements for Indian Languages for HTML 5.0
- Multilingual requirements for Voice Browser in Indian Languages
- Gap Analysis for implementation of Mobile Web best practices in Indian Languages
- Multilingual Linked Data for E-Governance data

The Internationalization Best Practices Would be foundation of the all the above initiatives
Thanks & Questions