

Roadmap for Localization & Language Technology Standards (Draft 1.0)

Inspired by Shri Ajeer Vidya, Chairman

Working Group- Localization & Language Technology Standards

Prepared By: 1. Shri M. D. Kulkarni & C-DAC Gist Team

2. Shri Kewal Krishan, Technical Director, NIC

Sr.No	Area	Current Issues/Status	Action/Destination Desired
1. OS Support			
1.1	OS Support under Windows, Linux, MAC OS	<p><u>Present Status :</u></p> <ul style="list-style-type: none"> In Windows 2000/XP -12/22 constitutionally recognized Indian Languages support is available. (Bangla, Gujarati, Hindi, Kannada, Konkani, Malayalam, Marathi, Punjabi, Sanskrit, Tamil, Telugu & Urdu). 14/22 under Windows-Vista added support for Oriya and Assamese. In RedHat Linux -9/22 constitutionally recognized Indian Languages support is available. (Bangla, Gujarati, Hindi, Marathi, Oriya, Punjabi, Malayalam, Tamil & Telugu). In MAC OS X-4/22 Gujarati, Hindi, Punjabi, Tamil constitutionally recognized Indian Languages Support is available 	<ul style="list-style-type: none"> In Windows OS : Bodo, Dogri, Kashmiri, Maithili, Manipuri, Nepali, Sindhi, Santhali Constitutionally recognized languages - support is not available In Red Hat Linux : Assamese, Bodo, Dogri, Kannada, Konkani, Kashmiri, Maithili, Manipuri, Nepali, Sindhi, Santhali, Sanskrit, Urdu Constitutionally recognized languages - support is not available In MAC OS X - 18/22 constitutionally recognized languages support is not available. <p>Requirement: all 22 constitutionally recognised languages support must be available in all major OS.</p>
1.2	Locales Data	Available for 12/22 languages under Windows, 9/22 languages under Linux in UTF-8 format.	<ul style="list-style-type: none"> Presently Locales data is insufficient and not accommodate Indian culture specific requirements. Detailed study to be undertaken.
1.3	Sorting	Presently sort order is on value.	<ul style="list-style-type: none"> Sort order for all Indian languages can be standardized and developers need to adhere to it.

*** For Bodo, Dogri, Maithili, Santhali and Kashmiri languages see Annexure - I

Roadmap for Localization & Language Technology Standards (Draft 1.0)

1.4	Encoding	<p>Unicode characters are almost complete to suffice the respective language requirements.</p> <ol style="list-style-type: none"> 1. Vedic Sanskrit code is being evolved. 2. Work for representing recently recognized constitutionally languages i.e. Bodo, Maithili, Santhali & Dogri are being initiated. 	<ol style="list-style-type: none"> 1. Constant interaction with Unicode for proper representative of Indian languages. 2. There has to be standards enforcement at State level. <p>*** For Bodo, Dogri, Maithili, Santhali and Kashmiri languages see Annexure - I</p>
1.5	<p>Inputting Mechanism</p> <ol style="list-style-type: none"> a) Keyboard Layouts b) Speech to Text c) Handwriting Recognition, Text OCR and Other inputting mechanisms 	<p><u>Present Status</u></p> <ol style="list-style-type: none"> a) Keyboard Layouts -Any Inputting method can be used in Unicode enabled OS. - INSCRIPT keyboard layout is available at OS level. b) Shrutlekhan-Rajbhasha is available for Hindi. c) Technology under development 	<ul style="list-style-type: none"> • Typewriter keyboard support as well as State level Language specific requirements (KGP keyboard layout for Kannada, TAM99 keyboard layout for Tamil) should be supported at operating system level. <p>Note: Output of any user specific keyboard layout must conform to Unicode current version.</p>
1.6	Rendering on PC in Application/Browser	For rendering Open Type fonts - rasterisation engine needs to be built in to the OS.	<ul style="list-style-type: none"> • Rasterisation engine is not being implemented for all the 22 scheduled Indian languages. Needs to be implemented by OS developers • Collation standards needs to be defined
1.7	Searching	<ul style="list-style-type: none"> • Character level search is available. • Contextual based search and intelligent search engines are not available 	Intelligent search engines required to be developed, however a huge amount of linguistic resources are required to get a fairly good accuracy.

***** For Bodo, Dogri, Maithili, Santhali and Kashmiri languages see Annexure - I**

2. Content Creation			
2.1	Content Creation Editors for Desktop & Web W3C :Markup languages Standard Generalized Markup Language (SGML) old std. Hypertext Markup Language (HTML) Extensible Markup Language (XML) Extensible Hypertext Markup Language (XHTML) XLIFF - XML Localization Interchange File Format TEI - Text Encoding Initiative	<ol style="list-style-type: none"> 1. Lot of tools is now available for content creation in Indian languages. 2. Products from C-DAC and open source tools such as Bharateeya OpenOffice can be used for the same 	
2.2	Browser Support	In IE6, IE7, FireFox, Netscape etc. - Indian Languages support is available.	
3. Resources & Tools			
3.1	Processing Resources : Spell Checker	Available for most of the Indian languages but need to be bettered.	<ol style="list-style-type: none"> 1. Enhanced version of spellcheckers needed 2. No attempt is made so far for building Grammar checker in Indian languages
3.2	Language Resources : Dictionaries, Glossary, Lexicon, Thesaurus, WordNet, Corpora: Text & Speech ISO : TermBase eXchange (TBX): ISO : Terminology Markup Framework (TMF) ISO : Lexical Resource Markup Framework (LRMF) EAGLES/ISLE: CES: Corpus Encoding Standards EAGLES/ISLE: XCES: XML based Corpus Encoding Standards EAGLES:MATE - Multilingual Annotation Tools Engineering	No Language Resources conforming to standards are available.	Needs to have a National initiative for development of Linguistic resources.
3.3	Machine Translation	<p>Research and Development in MT has been underway at several organizations in India.</p> <ol style="list-style-type: none"> i) English to Indian language MT Systems ii) Indian language to Indian language MT Systems iii) English has been the language of choice in the foreign language 	

Roadmap for Localization & Language Technology Standards (Draft 1.0)

		<p>category among MT R&D community in India. Efforts are concentrated around building MT systems for English-Hindi language pair. For Indian language to Indian language translation systems researches are focusing on developing MT systems for Hindi and other Indian languages.</p> <p>iv) University of Hyderabad has worked on an English-Kannada MT system, using the Universal Clause Structure Grammar (UCSG) formalism, invented there. This is essentially a transfer-based approach, and has been applied to the domain of government circulars.</p> <p>v) Some other organizations are also working in the area of Machine Translation such as IIT Kanpur (using Anglabharati approach), IIT Mumbai (using Universal Networking Language-UNL approach), Super Infosoft Pvt (developed Anuvadak system), IBM, Gurgaon (using Statistical approach), IIIT- Hyderabad and University of Hyderabad (developed Anusaraka - A Language Accessor).</p> <p>vi) CDAC, AAI Group has developed MT system (English to Hindi MT System) for Administrative, Finance, Agriculture and Small Scale Industry domains.</p>	
<p>3.4</p>	<p>Transliteration INSORT ISO</p>	<p>For few Indian languages Transliteration is available and is in use.</p> <p>While for Unicode based transliteration a separate add-on utility is required to be built.</p>	<p>For database translations region specific knowledge is required to build in the transliteration tools.</p>
<p>3.5</p>	<p>Database Support : Data Storage & Retrieval</p>	<ol style="list-style-type: none"> 1. If databases are Unicode complaint then there are no issues in regards with storage. Most of the databases now support Unicode. 2. However, there are issues in querying the data, searches & sorting in Indian languages 	<p>Database vendors needs to give support to Indian language query mechanism</p> <p>Or</p> <p>External API libraries may be developed for such purposes.</p>

*** For Bodo, Dogri, Maithili, Santhali and Kashmiri languages see Annexure - I

4. Search Engine Supporting Indian Languages (Google, Yahoo etc)			
	W3C	Presently character level search is available in all major search engines.	Needed to be developed as an plugin to the existing browsers as well as development of server based components for language search and appropriate intelligent crawler for index generation.
5. Localized Applications			
	<p>Open Office</p> <p>Works on Operating systems which has language support such as Windows XP, Linux.</p>	<ol style="list-style-type: none"> 1. C-DAC, GIST, Pune has already taken the work of localization of Bharateeya Open Office for all scheduled 22 Indian languages. 2. Currently localized versions for Tamil, Hindi & Telugu are released. 3. Localized versions for Kannada, Punjabi, Urdu, Oriya, Assamese, Bengali, Malayalam, Gujarati are ready and awaiting for release 4. Rest of the language localization is in progress. 	

- Since source code is available for Linux and it supports Unicode we can immediately start customizing it for e-Gov application need. Like Indix is already available for 12/22 Indian Languages.

Necessary changes for development of Open type fonts for Maithili, Santhali, Bodo & Dogri

This document is a report of the study done for the development of Open type font for remote languages. It demonstrates the necessary changes in the Devanagari script adapted by each of these languages in order to represent their tones and voice modifications. This is a result of an extensive study of these languages, their literature, script and culture.

The document is divided into 3 parts:

Part 1 is a summary of all the extra characters used by these languages and which do not fall at present in the codepage 900: Devanagari

In **Part 2**, the Input mechanism and storage issues are addressed, to ensure that these scripts are compatible with the existing code-page and that in storage there is no ambiguity or error of data representation.

In **Part 3** the display issues are addressed.

PART 1

Maithili, Santhali, Bodo & Dogri are using Devanagari script, modified to suit the specific requirements of the language. Some of the characters already exist in the Unicode page but have been allotted specific values.

This part shows in tabular format the additions to the existing Devanagari code-page and their respective use:

Character	Maithili	Santhali	Bodo	Dogri
' Latin apostrophe: (U0027)	--	--	To represent extended pronunciation (e.g. 1A)	To represent a high falling tone of previous short vowel as well as a syncopation marker (e.g. 1B)
◌̣ Devanagari Halant (U094D)	--	To indicate a half consonant (e.g. 1C)	--	Used frequently with Devanagari Ha (U0939) to represent a high falling tone of previous long vowel (e.g. 1D)
◌̣ Devanagari Nukta (U093C)	--	Used under Devanagari vowel sign AA (U093E) to represent a Samriddha swar (e.g. 1E)	--	--
◌̣ Devanagari sign Chandrabindu (U0901)	To represent nasalization used extensively with vowel sign E (U0947) (e.g. 1F)	--	--	--

1A

न'नख'र आरो न'खरआरी बेसादफोर गाबह'रो

1B

जि'या कन्नै द'ऊं वर्गे

1C

रेयाक् जामोक्

1D

पैह् ला ध्या साह्ब

1E

चितार हालियाक्

1F

मैथिली केँ जन भाषा

Of these additional characters, the only character which needs to be added to the Devanagari code-page is the apostrophe comma, since the other characters are already included in the DV code-page but have specific functions. This is more an input and storage issue.

It is strongly recommended that the apostrophe be included as a special character in the existing code-page of Devanagari Unicode.

A simplistic solution would be to use the Latin apostrophe: U0027. However this solution is to be rejected since both Dogri & Bodo are included in the scheduled list of languages and a user would be eventually permitted to create an IDN using these languages. Since IDN norms exclude mixing of code-pages, users of these languages will not be able to create IDNs where the apostrophe occurs. Given that the character is a length/tone marker in these languages and has a high incidence, it is imperative that it be included in the Devanagari Unicode set.

PART 2 : Key input and Storage

In this part issues pertaining to the input and storage of each specific character outlined above are laid out:

Character	Language	Key (INSCRIPT)	Storage
' Latin apostrophe: (U0027)	Dogri & Bodo	Shift + J	To be stored as apostrophe with value of Devanagari code page. At present will be mapped to ₹ (U0931) in storage
◌̣ Devanagari Halant (U094D)	Dogri & Santhali	Character + (Ctrl +d) + (Ctrl + Shift + 2)	Character + Halant (U094D)+ Zero width non joiner (U200C)
◌̣̣ Devanagari Nukta (U093C)	Santhali	Vowel sign AA +]	Devanagari vowel sign AA (093E) + Nukta (U093C)
◌̣̣̣ Devanagari sign Chandrabindu (U0901)	Maithili	Vowel sign E + (Shift + X)	Devanagari vowel sign E (U0947) + Chandrabindu (U0901)

Part 3 : Display issues

1. ' Latin apostrophe (U0027):

Current: The Latin apostrophe mark is placed above the 'Shirokeha', hence the difference between the use of Apostrophe as a punctuation mark and as a tone marker is not obvious.

Proposed: The apostrophe used as a tone marker should be aligned to the 'Shirokeha' to avoid the confusion.

2. ङ Devanagari Halant (U094D)

Ha followed by Halant should not join the preceding conjunct.

3. ङ Devanagari Nukta (U093C)

Current: Nukta not aligned below the Vowel sign AA.

Proposed: Nukta should be aligned below the Vowel sign AA.

4. ॅ Devanagari sign Chandrabindu (U0901)

Current (in CDAC fonts): Vowel sign E and Chandrabindu appear in their original size.

Proposed: The size of Chandrabindu should be reduced when it appears after the vowel sign E.

ISSUES PERTINENTES TO KASHMIRI:

- Often treated as a language of the Dardic family, Kashmiri is written using a modified Perso-Arabic script.
- Kashmiri has the highest number of vowels and therefore a complicated system of notation using diacritics and special vowel and palatalisation markers has been evolved.

Problem 1: Missing characters:

- The Ulta Pesh (ٲ) placed above the vowel sign: ٲ to indicate the lengthening of the vowel is still to be incorporated in Code Page 600
- Due representation has been made to Unicode by the TDIL and the NCPUL to introduce this sign used also in Urdu.
- Characters present in Codepage 600 but not specifically attributed to Kashmiri
 - Upper rounded sukun 06EB defined in Unicode as Arabic empty centre high stop.
 - Ulta Jazm 065A defined in Unicode as: Arabic Vowel sign: small v above. African Languages.
- Since these characters as per Unicode definition seem to have very little pertinence, they are not supported in fonts and in the case of 06EB, USP10.dll does not provide support, with the result that the Kashmiri script cannot be fully rendered.
- Unicode has to be persuaded to add that these two characters are part of the Kashmiri character set.

Kewal Krishan
Technical Director & Member Secretary
Localisation & Language Technology Standards
National Informatics Centre
New Delhi - 110003
E-Mail : kewal.krishan@nic.in
Ph : 011-24619860(O)
0-9810031413 (Cell)