# UNICODE STANDARD

For

# Indic Scripts

## UTC#94

## (UNICODE TECHNICAL COMMITTEE MEETING)

A Presentation by

**Dr. Om Vikas**

Senior Director & Head

Human Centred Computing Division

Government of India

Ministry of Communications & Information Technology

Department of Information Technology

omvikas@mit.gov.in

Tel/Fax : 91-11-2436 3076

## Contents

- National & International Language Scenario
- Language Technology Development in India
- Unicode Revision for Indic Scripts
- Unicode Conference in India

# Linguistic Scenario in India (1991 Census based)

| Language | Script | Population | Percent |
|---|---|---|---|
| Hindi | Devanagari | 33,72,72,114 | 41.6 |
| Bangla | Bengali | 6,95,95,738 | 8.6 |
| Telugu | Telugu | 6,60,17,615 | 8.1 |
| Marathi | Devanagari | 6,24,81,681 | 7.7 |
| Tamil | Tamil | 5,30,06,368 | 6.5 |
| Urdu | Urdu | 4,34,06,932 | 5.4 |
| Gujarati | Gujarati | 4,06,73,814 | 5.1 |
| Kannada | Kannada | 3,27,53,676 | 4.0 |
| Malayalam | Malayalam | 3,03,77,176 | 3.7 |
| Oriya | Oriya | 2,80,61,313 | 3.5 |
| Punjabi | Gurumukhi | 2,33,78,744 | 2.9 |
| Assamese | Assamese | 1,30,79,696 | 1.6 |
| Kashmiri | Urdu/Devanagari | 32,00,000 | 0.4 |
| Sindhi | Urdu/Devanagari | 21,22,848 | 0.3 |
| Nepali | Devanagari | 20,76,645 | 0.25 |
| Konkani | Devanagari | 17,60,607 | 0.20 |
| Manipuri | Manipuri | 12,70,216 | 0.15 |
| Sanskrit | Devanagari | 49,736 | 0.0006 |

# International Linguistic Scenario- (Hozumi Tanaka 1999)

## Language-wise

| Language | 2050 Population in Billion | 1996 Population in Billion |
|---|---|---|
| Chinese | 1.384 | 1.113 |
| Hindi/Urdu | 0.556 | 0.316 |
| English | 0.508 | 0.372 |
| Spanish | 0.486 | 0.304 |
| Arabic | 0.482 | 0.201 |
| Portuguese | 0.248 | 0.165 |
| Bengali | 0.229 | 0.125 |
| Russian | 0.132 | 0.155 |
| Japanese | 0.108 | 0.123 |
| German | 0.091 | 0.102 |
| Malay | 0.080 | 0.047 |
| French | 0.076 | 0.070 |

## Characteristics of Indian Languages:

- What You Speak Is What You Write (WYSIWYW)
- Script grammar describes transformation rules
- Relatively word-order-free
- Common phonetic based alphabet
- Common concept terms (from Sanskrit)

## *A B C Technology Development Phases*

Development of Language Technology in India may be categorized in three phases:

- 1971-1990 : **A**-Technology Phase

Focus was on **A**daptation Technologies; abstraction of requisite technological designs and competence building in R&D institutions.

- 1991-2000 : **B**-Technology Phase

Focus was on developing **B**asic Technologies- generic information processing tools, interface technologies and cross-compatibility conversion utilities. TDIL program was initiated.

- 2001-2010 : **C**-Technology Phase

Focus is on developing **C**reative Technologies in the context of convergence of computing, communication and content technologies. Collaborative technology development is being encouraged to realise.

# TDIL Vision 2010

**Vision statement**

Digital unite and knowledge for all.

**Mission statement**

Communicating without language barrier & moving up the knowledge chain.

**Objectives**

- To develop information processing tools to facilitate human machine interaction in Indian languages and to create and access multilingual knowledge resources/content.
- To promote the use of information processing tools for language studies and research.
- To consolidate technologies thus developed for Indian languages and integrate these to develop innovative user products and services.

# TDIL Mission - Major Initiatives

- **Knowledge Resources**
  (Parallel Corpora, Multilingual Libraries/Dictionaries, lexical resources)
- **Knowledge Tools**
  (Portals, Language Processing Tools, Translation Memory Tools)
- **Translation Support Systems**
  (Machine Translation, Multilingual Information Access, Cross Language Information Retrieval)
- **Human Machine Interface System**
  (OCR, Voice Recognition Systems, Text-to-Speech System)
- **Localization**
  (Adapting IT Tools and solutions in Indian Languages)
- **Language Technology Human Resource Development**
  (in NLP & Computational Linguistics)
- **Standardization**
  (ISCII, Unicode, XML, INSFOC, MPEG, Terminology, etc.)

**Long Term Goals**

- **Speech - to - Speech translation.**
- **Human Inspiring Systems**

# TDIL Programme - Achievements

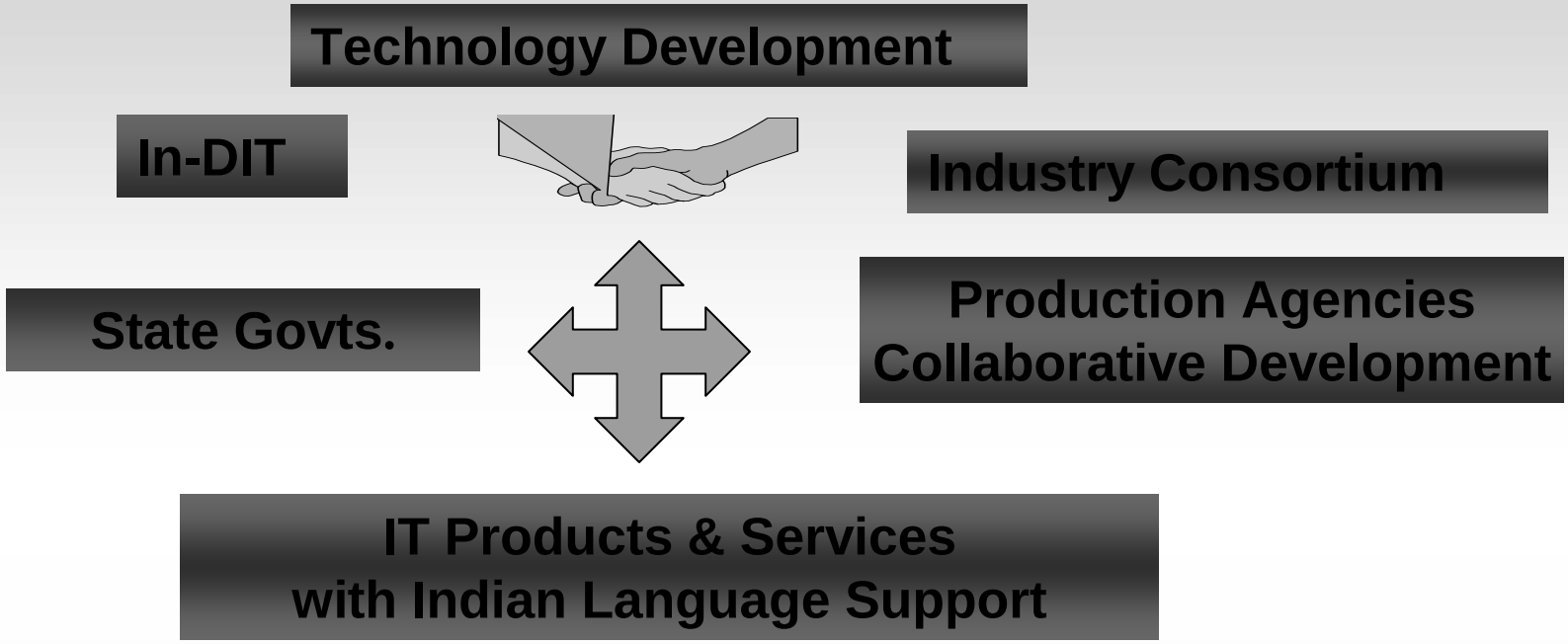
## Products and Services

- Machine Readable Corpora of Texts for Indian Languages
- Machine Translation Systems for English to Hindi
- Web-based Language Accessor from Indian languages to Hindi
- Internet Tools and Technologies for accessing web in Hindi
- PC based Speech Synthesis / Text-to Speech System
- OCR for Indian Scripts
- Word processors, Authoring systems and Applications packages
- Indian Language support on Linux

## Standardisation

- Standardisation of 8-bit ISCII, Keyboard layout.
- UNICODE for Indian Scripts
- Indian Standard Font Code
- Indian Script to Romanization & Vice versa Transliteration
- Indian Script to Indian Script Transliteration
- Lexware Formats

# Indian Language Technology Resource Centres




- To build repository of Indian Languages tools and products.
- Collaborative Developments in association with industry
- To develop niche technologies for providing IT localization solutions
- Technology dissemination through
  - IT localization clinics
  - Interaction with State Government for e-governance
  - Specialized training programs

## India's Efforts:

### Feedback on Present Unicode Standard for Indian Scripts:

- There are 18 constitutionally recognised languages in India.
- 10 Indic scripts are in use.
- Characteristics of Indian languages are common in general but different at some places.
- Department of IT, initiated the Survey / Discussions to get the feedback from State Governments, Experts and Industry to find out, if the existing Unicode Standard for Indic scripts are fulfilling the requirements of all the Indian languages.

## Consensus Building

• Department of Information Technology, Government of India, organised a number of discussion meetings of Experts/ Linguistic groups/ Industry to identify the solutions to overcome the deficiencies in the existing Unicode Standard for Indian scripts.

• It took pretty good time to reach to an agreement so that all the Indian languages are represented properly.

## Unicode Constraints

•difficult to change the existing Unicode Standard, so that a script can be properly represented in the Unicode Standards.

• Particularly the nomenclature of a character is very important issue. Once a character is encoded with wrong name, it always creates confusion among the user group.

• Take the case of Devanagari sign *Halant*, it is named as *Viram*, which is a punctuation mark in Devanagari. Though the annotation has been given with the character name, there should be some mechanism so that the correct name is popularised among the users and reference to wrong name is discouraged.

## Language Design Guides for Industry

• In consultation with the linguists/ experts the Languages Design Guides have been prepared for all the constitutionally recognized Indian Language to ease the Language Technology Industry and user groups.

• These guides provide in depth information about the languages their characteristics and other important issues.

• Detailed explanation about all the Indic scripts should be made available in the Unicode Standard.

# VEDIC SANSKRIT

## New Proposal for Vedic Sanskrit

Department of Information Technology has prepared a Coding scheme for encoding the Vedic Symbols and Characters in Unicode Standard. The draft of the proposed Vedic Code Set may be referred at:

**http://www.tdil.mit.gov.in/prop_uni/Vedic.pdf**

## Devanagari

Additions

| | | |
|---|---|---|
| 0979 | ग्‍ | DEVENAGARI LETTER Implosive GA |
| 097A | ज्‍ | DEVENAGARI LETTER Implosive JA |
| 097B | ड्‍ | DEVENAGARI LETTER Implosive DA |
| 097C | ब्‍ | DEVENAGARI LETTER Implosive BA |

These consonants are are exclusively used in Sindhi and have different Phonetic Value then GA,JA,DA,BA. Also these cannot be produced with the application of Anudatta and respective consonants.

## Devanagari

Additions

| | | |
|---|---|---|
| 0955 | ◌ॕ | DEVENAGARI ANUSWARA |
| 0956 | ◌ॖ | DEVANAGARI SIGN YAJURVEDIC ANUSWARA |
| 0957 | ◌ॗ | DEVANAGARI JIVHYAMULIYA |

These characters are commonly used in *Laukik Sanskrit shlokas* which are often used in Indian language text. Addition of these characters in the Devanagari code block will make the Devanagari code block complete in itself to meet the requirements of Hindi, Sanskrit, Marathi and other Devanagari based languages.

## Devanagari

Additions

| | | |
|---|---|---|
| 0972 | क्ष | DEVENAGARI LETTER KSHA |
| 0973 | ज्ञ | DEVENAGARI LETTER GNYA |
| 0974 | श्र | DEVENAGARI LETTER SHRA |
| 0975 | ॅ | DEVENAGARI SIGN REPH |
| 0971 | रु० | DEVENAGARI  CURRENCY SIGN |

The usage frequency of these characters in Devanagari based languages is very high and in some languages these are not considered to be the consonant conjuncts. The glyph of these characters are totally different from their base characters.

## Devanagari

Additions

0976 ऱ DEVENAGARI LETTER SHA

श (0936) ≡ ऱ (0976)

0977 ऴ DEVENAGARI LETTER LA

ल (0932) ≡ ऴ (0977)

0978 ॅ DEVENAGARI SIGN /SOFT RA/

These consonants are used in Marathi, and encoding these characters in Unicode will ease the transliteration among the Indian languages. They have same phonetic value but different shapes when used in Marathi language.

# Devanagari

## Change in Annotation/ Explanation

| | Present Explanation | New Explanation |
|---|---|---|
| 094D | DEVENAGARI SIGN VIRAM<br>= halant<br>Suppresses inherent Vowel | DEVENAGARI SIGN HALANT<br>To take out hidden vowel sound of the consonant.<br>Omission of Vowel /A/ |
| 0964 | DEVENAGARI DANDA<br>= Phrase separator | DEVENAGARI SIGN POORNA VIRAM |
| 0965 | DEVENAGARI DOUBLE DANDA | DEVENAGARI SIGN DEERGH VIRAM |

## Devanagari

Change in Glyph

| | Present | Proposed | |
|---|---|---|---|
| 0929 | ऩ | ऩ | DEVANAGARI LETTER NNNA<br>For transcribing Dravidian alveolar n<br>Fricative Consonant used in Marathi |
| 0931 | ऱ | ऱ | DEVANAGARI LETTER RRA<br>For transcribing Dravidian alveolar r<br>Half form is represented as “Eyelash RA”<br>Fricative Consonant used in Marathi |
| 0934 | ऴ | ऴ | DEVANAGARI LETTER LLLA<br>For transcribing Dravidian alveolar l<br>Fricative Consonant used in Marathi |

The Glyph of these characters need to be changed because the sound associated with these consonants are different from the consonants used with Nukta.

## Bengali

Additions

| | | |
|---|---|---|
| 09B1 | ৰ | BENGALI LETTER RA WITH MIDDLE DIAGONAL (Used in Assamese) |
| 09B5 | ৱ | BENGALI LETTER WA WITH LOWER DIAGONAL (Used in Manipuri & Assamese) |
| 09BD | ঽ | BENGALI SIGN AVAGRAH |
| 09FB | ◌্য | BENGALI LETTER YA FALLA |
| 09FC | ৎ | BENGALI LETTER KHANDA TA |
| 09FD | ক্ষ | BENGALI LETTER KSHA (Used in Assamese) |

- BENGALI LETTER YA-FALLA and BENGALI LETTER KHANDA TA can be produced with application of Hasant Sign, but the ligatures so obtained will depend upon the context and application where these consonants are being used.

- The letter KSHA is not treated as the consonant conjunct in Manipuri and Assamese languages hence to ease the transliteration it is desired that this consonant is encoded separately in the Bengali code block.

## Gujarati – Additions

| | | |
|---|---|---|
| 0A8C | ઌ | GUJARATI LETTER VOCALIC L |
| 0AE1 | ૡ | GUJARATI LETTER VOCALIC LL |
| 0AE2 | ૢ | GUJARATI VOWEL SIGN VOCALIC L |
| 0AE3 | ૣ | GUJARATI VOWEL SIGN VOCALIC LL (Used with Sanskrit text) |
| 0 AF6 | ૺ | GUJARATI FRACTIONAL NUMERAL /PA/ 1/4 |
| 0 AF7 | ૻ | GUJARATI FRACTIONAL NUMERAL /Addho/ 1/2 |
| 0 AF8 | ૼ | GUJARATI FRACTIONAL NUMERAL /Pono/ 3/4 |

- The fractional numerals and diacritic marks are very commonly used in Gujarati. Hence these characters should be added in the Gujarati code block.

- The Glyphs of characters at code point 0A8C and 0AE1 as shown in Unicode 4.0 Beta are not correct and need to be corrected.

## Gurmukhi - Additions

0A4E ੍ GURMUKHI PARI – RA (Consonant modifier)

0A4F ੏ GURMUKHI PARI – HA (Consonant modifier)

0A76 ੶ GURMUKHI PARI VA [Halant ੍ (0A4D) ਵ (0A35)]

0A77 ੷ GURMUKHI PARI YA [Halant ੍ (0A4D) ਯ (0A2F)]

0A78 ੸ GURMUKHI Half Y-YA [ ਯ (0A2F) Halant ੍ (0A4D)]

0A79 ੹ GURMUKHI PARI GU [Halant ੍ (0A4D) ਗ (0A17)]

0A7A ੺ GURMUKHI PARI CA [Halant ੍ (0A4D) ਚ (0A1A)]

0A7B ੻ GURMUKHI PARI TA [Halant ੍ (0A4D) ਤ (0A24)]

0A7C ੼ GURMUKHI PARI NA [Halant ੍ (0A4D) ਨ (0A2B)]

## Gurmukhi

Additions

| | | |
|---|---|---|
| 0A50 | ੴ | GURMUKHI EK ONKAR (God is one) |
| 0A74 | ੵ | GURMUKHI SIGN KHANDA |
| 0A01 | ਁ | GURMUKHI TIPPI (Nasalization) |
| 0A03 | ਃ | GURMUKHI VISARG |
| 0A3B | ਼ | GURMUKHI ADDAK (Doubles following consonant) |
| 0A7D | ੰ | GURMUKHI VOWEL SIGN BIHARI BINDI |
| 0A7E | ੱ | GURMUKHI VOWEL SIGN AAN (KANA BINDI) |

# KANNADA

## Addition

0CBB ◌ಾ KANNADA VOWEL SIGN /A/

0CBC ◌಼ KANNADA SIGN NUKTA

0CBD ಽ KANNADA SIGN AVAGRAH

0CF5 ೱ KANNADA SIGN REPH

0CF9 ◌೹ KANNADA DIACRITIC SIGN DEERGHA SWARITHA

# Kannada

## Change in Annotation/Explanation

| | | |
|---|---|---|
| 0CC0 | KANNADA VOWEL SIGN II | KANNADA VOWEL SIGN II |
| | ≡ 0CBF ◌ೀ 0CD5 ◌ೕ | |
| 0CC7 | KANNADA VOWEL SIGN EE | KANNADA VOWEL SIGN EE |
| | ≡ 0CC6 ◌ೆ 0CD5 ◌ೕ | |
| 0CC8 | KANNADA VOWEL SIGN AI | KANNADA VOWEL SIGN AI |
| | ≡ 0CC6 ◌ೆ 0CD6 ◌ೖ | |
| 0CCA | KANNADA VOWEL SIGN O | KANNADA VOWEL SIGN O |
| | ≡ 0CC6 ◌ೆ 0CC2 ◌ೂ | |
| 0CCB | KANNADA VOWEL SIGN OO | KANNADA VOWEL SIGN OO |
| | ≡ 0CCA ◌ೊ 0CD5 ◌ೕ | |

## MALAYALAM

### Additions

| | | |
|---|---|---|
| 0D7A | ൺ | MALAYALAM LETTER NN |
| 0D7B | ൻ | MALAYALAM LETTER N |
| 0D7C | ർ | MALAYALAM LETTER RR |
| 0D7D | ൽ | MALAYALAM LETTER L |
| 0D7E | ൾ | MALAYALAM LETTER LL |

- There are five consonants in the Malayalam known as *Chillu,* are pure consonants i.e. like other consonants these consonants do not have a inherent vowel /a/.
- The pure consonants are also represented in half form of the characters but this half form go into constitution of consonant conjuncts.
- *Chillu* are the only consonants in Malayalam, which are represented independently irrespective that these are pure consonants.

## Malayalam

- The glyph Malayalam vowel sign AU (0D4C) should be corrected.
- The code point 0D4D ( Malayalam Sign Viram) should be read as Malayalam Sign CHANDRAKKALA.

# Oriya

## Change in Glyph

| | Present | New Glyph | |
|---|---|---|---|
| 0B48 | ◌ୈ | ◌ୈ | ORIYA VOWEL SIGN AI |
| 0B4C | ◌ୌ | ◌ୌ | ORIYA VOWEL SIGN AU |
| 0B66 | ୦ | ୦ | ORIYA DIGIT ZERO |

## Oriya

## Additions

0B35 ଵ ORIYA LETTER WA

0B44 ◌ୄ ORIYA VOWEL SIGN VOCALIC RR

- In the Oriya Script there are BA and WA.
- There is no VA in Oriya Script.

## Change in Explanation

| | Present Explanation | New Explanation |
|---|---|---|
| 0B2C ବ | ORIYA LETTER BA<br>=Oriya va, wa | ORIYA LETTER BA |
| 0B5C ଡ଼ | ORIYA LETTER RRA | ORIYA LETTER DDA |
| | 0B21 ଡ 0B3C ◌଼ | =0B21 ଡ 0B3C ◌଼ |
| 0B5D ଢ଼ | ORIYA LETTER RHA | ORIYA LETTER DDHA |
| | 0B22 ଢ 0B3C ◌଼ | =0B22 ଢ 0B3C ◌଼ |

# Tamil

## Change in Glyph

0B83 ◌ஃ ஃ TAMIL LETTER AYTHAM

## Change in Annotations/ Explanation

| | | Present Explanation | New Explanation |
|---|---|---|---|
| 0B82 | ◌ஂ | TAMIL SIGN ANUSVARA | TAMIL SIGN ANUSVARA<br>Not used in Tamil |
| 0B83 | ◌ஃ | TAMIL SIGN VISARGA | TAMIL LETTER AYTHAM |
| 0BCA | ொ | TAMIL VOWEL SIGN O<br>Pieces on both sides of the consonant<br>≡ 0BC6 ெ 0BBE ◌ா | TAMIL VOWEL SIGN O |
| 0BCB | ோ | TAMIL VOWEL SIGN OO<br>Pieces on both sides of the consonant<br>≡ 0BC7 ே 0BBE ◌ா | TAMIL VOWEL SIGN OO |
| 0BCC | ௌ | TAMIL VOWEL SIGN AU<br>Pieces on both sides of the consonant<br>≡ 0BC6 ெ 0BD7 ◌ௗ | TAMIL VOWEL SIGN AU |

# TELUGU

## Additions

0C3C ◌ఀ TELUGU SIGN NUKTA

Placed at bottom left corner of the letter

0C3D ఀ TELUGU SIGN AVAGRAH

There is a strong need of adding (Telugu) diacritic marks in the Unicode Standard for Telugu, as there is no independent characters in the scripts for representation of sounds borrowed from other languages. These characters can be produced with the use of Nukta and respective consonants.

## **ARABIC -** Urdu, Sindhi and Kashmiri

- ARABIC LETTER ULTA PESH (ۭ), ARABIC LETTER JAZM (ۡ), ARABIC LETTER KHARI ZER (ۭ), ARABIC LETTER HAMZA (ٔ)(Diacritical) and ARABIC LETTER BAT (ٕ) need to be added in the Arabic code set.
- The code points 068E, 06B2 and 06B4 are not used in Sindhi hence an annotation should be added as “Not used in Sindhi”.
- The code point 0690 is not used in Urdu so annotation should be added “Not used in Urdu”.
- The shape of code points 0664, 0665,0666 and 0667 should be modified for proper representation of Sindhi-Arabic. The correct shape for these code points respectively are:

( ۴ ),( ۵ ),( ۶ ) and ( ۷ )

## General Observations

- In the Unicode code set for all the Indian script the *HAL or HALANT* sign is named as *VIRAM,* which is wrong.

- *Viram* is a punctuation mark in Devanagari script, which is also mentioned as *DANDA*. The *DEERGH VIRAM* is named as Double DANDA, "*Double"* is not an Indian word.

- Except for Devanagari no other code set is provided with Punctuation Marks *PURNA VIRAM* and *DEERGH VIRAM.* These signs need to be added in the code set for all the languages, as they need to be recognised for the purpose of proper rendering.

- The diacritic marks also extensively used in the Indian script text. The diacritic marks need to be added in the code set of Indian scripts.

# ‘Invisible Character’ for Indic Scripts

There is a need to add INVISIBLE character in the Indic Script code blocks. There are requirements to display the vowel sign independently in the text material.

# Pipe-Line Table for Unicode 4.0.0

## - Reference to Indic Scripts

- The representative Glyphs for some characters proposed to be included in Unicode 4.0.0 need to be corrected.
- These characters are Gujarati Letters Vocalic L and Vocalic LL
- Oriya Letter VA and WA both have been included for Unicode-4.0.0. The letter VA is not used in Oriya.
- The Glyph of Bengali Sign Avagraha is also to be corrected as ৲

## Helpline

Government of India, Department of Information Technology initiated the TDIL Programme, to address various issues related to use of Indian Languages in Information Technology. The website of the tdil programme is: **http://www.tdil.mit.gov.in**

and an email service **info@tdil.mit.gov.in** is to answer various queries of the user.

## Suggestions

### Unicode International Conference / Workshop

India is a multilingual & multiscript developing country. It is proposed that Unicode International Conference be organised in India to benefit the Multilingual Software Industry in India and the neighboring countries. This will promote multilingual content creation and cross lingual Information access to bridge the sprawling digital divide.

Government, Academia, Industry **together** to play **globally** and to serve **locally** in multilingual computing.

ધન્યવાદ

ਧਨਯਵਾਦ

ધન્યવાદ

# Thank You