A Proposal from Tamil Nadu Government for Tamil Unicode:

Presented by

Dr. M. Ponnavaikko
Former Director, Tamil Virtual University, & Vice-Chairman, Task Force on TACE-16
Director (Research & Virtual Education)
SRM University

Representing Tamil Nadu Government.
A Proposal from Tamil Nadu Government for Tamil Unicode

and by

Mr. Mani M. Manivannan
Director of Engineering, Symantec Corporation
Mountain View, CA

Founding Exec. Committee Member, INFITT,

Member, Task Force on TACE-16

Chairman, Tamil Internet 2002 Conference, Foster City, CA.

Founder, TSCII.ORG.
Agenda

- TACE-16 Task Force and its Mission
- Tamil language and the Nature of its Script
- Current Tamil Encodings and their Limitations
- Efforts to develop efficient, true 16-bit encoding
- TACE-16 Encoding and its merits
- Presentation, Testing and Reviews of TACE-16
- Proposal to Unicode
TACE-16 Task Force

- Constituted by Government of Tamil Nadu
- Consists of experts from academia and industry from Tamil Nadu, Government of India and from the Tamil Diaspora
- To evaluate, disseminate and recommend to declare TACE-16 as a Tamil encoding standard for IT applications in Tamil
- To present TACE-16 to Unicode Consortium for incorporation into the Unicode standard
What are the IT needs?

- 65 million Tamils in India, 80 million worldwide
- Millions of petitions, commercial transaction registrations, birth/death records, are generated in Tamil language every year.
- The TN government is in the process of digitizing its billions of records as a precursor to the e-governance projects
TN Government’s Tamil IT initiatives - 1

- TamilNet ’99 conference
  - 8 bit glyph encoding standards (TAM/TAB)
  - Keyboard standardization (phonetic/typewriter)
  - Evolving 16-bit character encoding for Tamil for incorporation into Indian national and Unicode standards
    - Became an Associate member of Unicode Consortium
- Formation of Tamil Virtual University
- Initiative to form INFITT
Developed an efficient, true 16-bit all character encoding – called TUNE. Tested on various platforms and applications.

Presented the encoding at various Tamil Internet conferences held around the world.

Discussed the encoding in various fora including INFITT.

Placed TUNE in the Unicode Private Use Area at the suggestion of Unicode Consortium and sought and reviewed user community feedback.
TN Government’s Tamil IT initiatives - 3

- Held a conference in September ‘06 to review TUNE and incorporated feedback to develop TANE
- Tested on several platforms and applications to develop TACE-16
- Funding development of tools and drivers to support TACE-16 for free distribution
- Became a voting Institutional Member of Unicode Consortium to present TACE-16
- Sought and received support from Government of India
On Tamil Language

- Recognized as one of the Classical Languages of the World
- At least 2500 years of Inscriptional records
- 2000+ years of unbroken literary history
- Tolkappiyam, an ancient grammar (2000+ years old) – still governing the language
- Conservative Language – preserves continuity
- People passionate about language
Nature of Tamil Script

- Alpha syllabic writing system
- Includes Vowels, Consonants and Vowel-Consonants – all graphically represented as **SINGLE LETTERS** (Tolkappiyam, Elu. 17-18).
- “The nature of the consonant is to be provided with a dot (pulli).” (Tolkappiyam, Elu. 15-17).
- Script shape has changed over centuries but the syllabic characters and sounds remain the same
Tamil Scripts

- Tamil Language has 247 Characters

Vowels (12)

Aytham(1)
Tamil Scripts

Consonants (18)

Nature of consonants is to be provided with a dot. The short e and short o are also of the same nature. Tol. Elu. 15-17
# Uyir-Mey Characters *(Vowel Consonants)*

<table>
<thead>
<tr>
<th>த்</th>
<th>த்</th>
<th>ழ்</th>
<th>ழ்</th>
<th>ஸ்</th>
<th>ஸ்</th>
<th>஽்</th>
<th>஽்</th>
<th>ஜ்</th>
<th>ஜ்</th>
</tr>
</thead>
<tbody>
<tr>
<td>ம்</td>
<td>ம்</td>
<td>க்</td>
<td>க்</td>
<td>ர்</td>
<td>ர்</td>
<td>ன்</td>
<td>ன்</td>
<td>ள்</td>
<td>ள்</td>
</tr>
<tr>
<td>ல்</td>
<td>ல்</td>
<td>ள்</td>
<td>ள்</td>
<td>ழ்</td>
<td>ழ்</td>
<td>஽்</td>
<td>஽்</td>
<td>ஜ்</td>
<td>ஜ்</td>
</tr>
<tr>
<td>கௌ</td>
<td>கௌ</td>
<td>கௌ</td>
<td>கௌ</td>
<td>ரௌ</td>
<td>ரௌ</td>
<td>னௌ</td>
<td>னௌ</td>
<td>ளௌ</td>
<td>ளௌ</td>
</tr>
<tr>
<td>ப்ப</td>
<td>ப்ப</td>
<td>ப்ப</td>
<td>ப்ப</td>
<td>ர்ப்</td>
<td>ர்ப்</td>
<td>ன்ப்</td>
<td>ன்ப்</td>
<td>ள்ப்</td>
<td>ள்ப்</td>
</tr>
<tr>
<td>ப்ப</td>
<td>ப்ப</td>
<td>ப்ப</td>
<td>ப்ப</td>
<td>ர்ப்</td>
<td>ர்ப்</td>
<td>ன்ப்</td>
<td>ன்ப்</td>
<td>ள்ப்</td>
<td>ள்ப்</td>
</tr>
<tr>
<td>க்</td>
<td>க்</td>
<td>ர்</td>
<td>ர்</td>
<td>ன்</td>
<td>ன்</td>
<td>ள்</td>
<td>ள்</td>
<td>ஜ்</td>
<td>ஜ்</td>
</tr>
<tr>
<td>கௌ</td>
<td>கௌ</td>
<td>ரௌ</td>
<td>ரௌ</td>
<td>னௌ</td>
<td>னௌ</td>
<td>ளௌ</td>
<td>ளௌ</td>
<td>ளௌ</td>
<td>ளௌ</td>
</tr>
</tbody>
</table>

May 2007 Tamil Unicode Issues
Every Tamil child has been learning Tamil character set as this table for at least 2000 years. The character shapes may have changed over the centuries. But the characters and sound have remained the same. This is important. These are not glyphs, not ligatures, not compound characters. But are simple characters just like A, B, C, D are characters to English speaking children. ka, kA, ki, kI, are characters to Tamil children. This is the basis for Tamil All Character Encoding initiative.

SRM, 5/15/2007
Nature of Tamil Vowel-Consonants

- Every Tamil child has been learning Tamil character set as in the previous table for several centuries.
- Uyir-meys are not glyphs, not ligatures, not conjunct characters.
- Uyir-meys are simple characters just like A, B, C, D are characters to English speaking children.
- ka, kA, ki, kl, etc., are characters to Tamils.
- This is the basis for the development of Tamil All Character Encoding scheme.
**Grantha Letters**

To represent Sanskrit borrowals

<table>
<thead>
<tr>
<th>சே கே செ கெ சே கே</th>
<th>கே கெ சே கே கே கே</th>
<th>சே கே கெ சே கே</th>
<th>கே கே சே கே</th>
<th>கே கே சே கே</th>
</tr>
</thead>
<tbody>
<tr>
<td>மே கே மே கே மே கே</td>
<td>கே கே மே கே</td>
<td>மே கே கே மே</td>
<td>மே கே கே மே</td>
<td>மே கே கே மே</td>
</tr>
<tr>
<td>பே கே பே கே பே கே</td>
<td>கே கே பே கே</td>
<td>பே கே கே பே</td>
<td>பே கே கே பே</td>
<td>பே கே கே பே</td>
</tr>
<tr>
<td>தே கே தே கே தே கே</td>
<td>கே கே தே கே</td>
<td>தே கே கே தே</td>
<td>தே கே கே தே</td>
<td>தே கே கே தே</td>
</tr>
<tr>
<td>நே கே நே கே நே கே</td>
<td>கே கே நே கே</td>
<td>நே கே கே நே</td>
<td>நே கே கே நே</td>
<td>நே கே கே நே</td>
</tr>
<tr>
<td>பே கே பே கே பே கே</td>
<td>கே கே பே கே</td>
<td>பே கே கே பே</td>
<td>பே கே கே பே</td>
<td>பே கே கே பே</td>
</tr>
</tbody>
</table>
### Tamil Scripts

<table>
<thead>
<tr>
<th>Description</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total characters in Tamil including Grantha letters</td>
<td>325</td>
</tr>
<tr>
<td>Tamil Numerals</td>
<td>13</td>
</tr>
<tr>
<td>Special Characters</td>
<td>9</td>
</tr>
<tr>
<td><strong>Total code points required</strong></td>
<td>347</td>
</tr>
</tbody>
</table>
Tamil Scripts – Frequency Analysis

Usage of Tamil characters in plain text:

Vowel Consonants (uyir-meys) : 64 – 70%
Vowels (uyir) : 5 – 6%
Consonants (meys) : 25 – 30%

Breaking high frequency letters into glyphs is highly inefficient
தமிழ் குறிகள் பயன்படுத்தும் பட்டியலின் பற்றிய சொல்லிபெயர் பயன்படுத்தும் பட்டியலின் பற்றிய சொல்லிபெயர். கலையான சட்டங்களின் பயன்படுத்தும் பட்டியலின் பற்றிய சொல்லிபெயர். பயன்படுத்தும் பட்டியலின் பற்றிய சொல்லிபெயர். கலையான சட்டங்களின் பயன்படுத்தும் பட்டியலின் பற்றிய சொல்லிபெயர்.
Current Tamil Encodings

- ISCII – 7 bit
- TSCII/TAB – 7 bit
- TAM – 8 bit
- Unicode – 7 bit
- Proprietary encodings – 7/8 bit
Limitations of Current Encodings

- 7/8 bit – insufficient to represent all Tamil characters
- Hinders Natural Language Processing including parsing, searching, sorting, etc.
- Unnatural for Speech to Text/Text to Speech
- Inefficient to store, transmit and retrieve
- Complex processing - hinders software development
- Needs a rendering engine even for plain text
- Needs “normalization” for string comparison
Unicode Design Goals

Unicode Standard is designed to be

- Universal:

  The repertoire must be large enough to encompass all characters that are likely to be used in general text interchange, including those in major international, national, and industry character sets.
Unicode Design Goals

Unicode Standard is designed to be

- **Efficient**: Plain text is simple to parse; software does not have to maintain state or look for special escape sequences and characters synchronization from any point in a character stream is quick and unambiguous. A fixed character code allows for efficient sorting, searching, display and editing of text.
Unicode Design Goals

Unicode Standard is designed to be

- **Unambiguous**: Any given Unicode point always represents the same character
Unicode Tamil Encoding

• 16 bit space – 64,536 code points available.

• Based on 7-bit ISCII.

• Uses only only 128 code point block and that too is mostly empty.

• Encodes glyphs which have no sound and are not characters in Tamil.
Violation of Unicode principles in the Present Unicode Tamil Encoding

- All the characters of Tamil are not encoded as per the Universal principle of Unicode
  - Only 10% of the Tamil Characters are provided code space in the Present Unicode Tamil.
  - 90% of the Tamil Characters that are used in general text interchange are not provided code space.
  - These 90% of the Tamil Characters are the Vowel Consonants. Of these Vowel Consonants only following vowel consonants are encoded

akkhilladhu...
Violation of Unicode principles in the Present Unicode Tamil Encoding

- The other vowel consonants need to be rendered using the following Vowel Consonants and the vowel signs encoded in the standard through a specially designed Rendering Engine.

ஏ, இ, .........., ஐ
Violation of Unicode principles in the Present Unicode Tamil Encoding

- There are two methods of rendering the following Vowel Consonants

![Vowel Consonants Example]

This leads to ambiguity in rendering characters
Rendering of Vowel Consonants

- Code points: 0B9A (ṉ) + 0BCA (ெ்) → Character loon
- Code points: 0B9A (ṉ) + 0BC6 (ெ்) + 0BBE (ெ்) → Character łoń

- Level II encoding, Complex Character Set, Rendering Engine has to shape the character
- Same Character can be formed by two different sets of code points leading to ambiguity (canonical equivalence!)
Violation of Unicode principles in the Present Unicode Tamil Encoding

- The Present Unicode is not efficient for parsing.

Counting the letters in the name

மணிவண்ணன்

- Even a Tamil child in primary school can say that this name has SIX letters
- According to Unicode this name has Nine characters:

மண◌ ி வண◌் ணன◌ ்

- To properly count the letters in this name, someone has to write a complicated program, worth to present a technical paper on this in a Tamil computing conference!
- There is a lot of such problems in complex encoding like this.
Violation of Unicode principles in the Present Unicode Tamil Encoding

- The present Unicode Tamil is not efficient for sorting, searching and natural language processing
Unicode Design Goals

<table>
<thead>
<tr>
<th>Investigation Type</th>
<th>SCHEME 1 (Unicode 3.0)</th>
<th>SCHEME 2 (Consonant-Vowel)</th>
<th>Scheme 3 (All Character)</th>
</tr>
</thead>
<tbody>
<tr>
<td>File Size</td>
<td>147</td>
<td>142</td>
<td>100</td>
</tr>
<tr>
<td>Display time</td>
<td>2,500</td>
<td>2,875</td>
<td>100</td>
</tr>
<tr>
<td>File transfer time</td>
<td>147</td>
<td>142</td>
<td>100</td>
</tr>
<tr>
<td>Find &amp; replace time</td>
<td>270</td>
<td>257</td>
<td>100</td>
</tr>
</tbody>
</table>

1. Data storage, retrieval and display parameters

2. Database related parameters

3. Morphological analysis parameters

May 2007 Tamil Unicode Issues
Violation of Unicode principles in the Present Unicode Tamil Encoding

Unicode is not supported in many platforms

( a few user comments are given below)

- Only a handful of applications support Tamil Unicode
- Vendors not keen to enable to handle 'complex Indic scripts'
- Documentation for implementing Indic Scripts is very poor.
  - Implementers don't have detailed knowledge of the script
  - Need to depend on 'language experts' who often disagree
- Implementations of many of the Indic scripts have serious errors.
- Fonts are "very ugly" and are few
- Vendors are primarily catering to the Government market.
Problems of current Tamil Unicode in Display
More examples for display problems

- பாகாத்தை தீராம் புராத்தியில்லா சோட்டம்

- போன்ற போர்கள் பாட்ட.

- குறு காட்சியின் ப்ள்ளோட்டியை போக்கு போக்கிறது பாகாத்தை சிட்டில் உள்ளது பாகாத்தை

- தமிழர்கள் ஆறு போர்கள் போக்காக்கின

- தமிழாக் பயிர் தமிழ் போர்கள்
More Tamil Unicode examples

"வீட்டின் முழு உரிமையும் திருமதி இராஜாவுக்கு ஆகாடுத்து விடுகிறது என்று வெளியேறி வழங்கும்"

In some software, the grantha letters ஜ, ஷ, ஹ, are corrupted and the above becomes

"வீட்டின் முழு உரிமையும் திருமதி இராஷாவுக்கு ஆகாடுத்து விடுகிறது என்று வெளியேறி வழங்கும்"

Will the court give the property to திருமதி இராஜா or திருமதி இராஷா?

May 2007 Tamil Unicode Issues 35
Violation of Unicode principles in the Present Unicode Tamil Encoding

- The Unicode standard encodes characters, not glyphs

- Unicode Tamil standard includes the following vowel signs

Are they characters or glyphs?

TACE is not in violation of Unicode Character Encoding Model – it conforms to it, unlike present Tamil Unicode.
Violation of Unicode principles in the Present Unicode Tamil Encoding

- Unicode Tamil includes the vowel consonants

  க த ........

- The following Tamil scripts

  கா இமா ........... கா

  கா இமா ........... கா

  கா இமா ........... கா

  கா இமா ........... கா

are also Vowel – Consonants. But they are not encoded
Violation of Unicode principles in the Present Unicode Tamil Encoding

The Vowel – Consonants are not glyphs. They are characters, designed as

Consonant + Vowel = Vowel Consonants

e.g.

\[ \ddot{\text{ஞ}} + \text{ந} = \text{ஆ} \]
\[ \ddot{\text{ங}} + \text{த} = \text{திற} \]

Unicode provides for rendering

Which has no meaning in Tamil. This type of rendering does not help simple character parsing as in a compiler or an interpreter.
TN Govt. proposal for incorporating TACE – 16 in the Unicode BMP space

The Character set proposed:

### 16-bit Tamil All Character Encoding (TACE_16)

<table>
<thead>
<tr>
<th>xy0</th>
<th>xy1</th>
<th>xy2</th>
<th>xy3</th>
<th>xy4</th>
<th>xy5</th>
<th>xy6</th>
<th>xy7</th>
<th>xy8</th>
<th>xy9</th>
<th>xyA</th>
<th>xyB</th>
</tr>
</thead>
</table>
| 0   | க   | ங   | ன   | ர ந   | ல வ ம ச ஂ ல ஒ ப க ம ட ந ப ா த ம த க ப க ம ப ா த ம த க ப ओ ज़ य क र ज ल व भ न ः त न ब घ ञ झ घ ध न श ल र ऊ च न द भ न ः त न ब घ ञ झ घ ध न श ल र ऊ च न द भ न ः त न ब घ ञ झ घ ध न श ल र ऊ च न द भ न ः त न ब घ ञ झ घ ध न श ल र ऊ च न द भ न ः त न ब घ ञ झ घ ध न श ल र ऊ च न द भ न ः त न ब घ ञ झ घ ध न श ल र ऊ च न द भ न ः त न ब घ ञ झ घ ध न श ल र ऊ च न द भ न ः त न ब घ ञ झ घ ध न श ल र ऊ च न द भ न ः त न ब घ ञ झ घ ध न श ल र ऊ च न द भ न ः त न ब घ ञ झ घ ध न श ल र ऊ च न द भ न ः त न ब घ ञ झ घ ध न श ल र ऊ च न द भ न ः त न ब घ ञ झ घ ध न श ल र ऊ च न द भ न ः त न ब घ ञ झ घ ध न श ल र ऊ च न द भ न ः त न ब घ ञ झ घ ध न श ल र ऊ च न द भ न ः त न ब घ ञ झ घ ध न श ल र ऊ च न द भ न ः त न ब घ ञ झ घ ध न श ल र ऊ च न द भ न ः त न ब घ ञ झ घ ध न श ल र ऊ च न द भ न ः त न ब घ ञ झ घ ध न श ल र ऊ च न द भ न ः त न ब घ ञ झ घ ध न श ल र ऊ च न द भ न ः त न ब घ ञ झ घ ध न श ल र ऊ च न द भ न ः त न ब घ ञ झ घ ध न श ल र ऊ च न द भ न ः त न ब घ ञ झ घ ध न श ल र ऊ च न द भ न ः त न ब घ ञ झ घ ध न श ल र ऊ च न द भ न ः त न ब घ ञ झ घ ध न श ल र ऊ च न द भ न ः त न ब घ ञ झ घ ध न श ल र ऊ च न द भ न ः त न ब घ ञ झ घ ध न श ल र ऊ च न द भ न ः त न ब घ ञ झ घ ध न श ल र ऊ च न द भ न ः त न ब घ ञ झ घ ध न श ल र ऊ च न द भ न ः त न ब घ ञ झ घ ध न श ल र ऊ च न द भ न ः त न ब घ ञ झ घ ध न श ल र ऊ च न द भ न ः त न ब घ ञ झ घ ध न श ल र ऊ च न द भ न ः त न ब घ ञ झ घ ध न श ल र ऊ च न द भ न ः त न ब घ ञ झ घ ध न श ल र ऊ च न द भ न ः त न ब घ ञ झ घ ध न श ल र ऊ च न द भ न ः त न ब घ ञ झ घ ध न श ल र ऊ च न द भ न ः त न ब घ ञ झ घ ध न श ल र ऊ च न द भ न ः त न ब घ ञ झ घ ध न श ल र ऊ च न द भ न ः त न ब घ ञ झ घ ध न श ल र ऊ च न द भ न ः त न ब घ ञ झ घ ध न श ल र ऊ च न द भ न ः त न ब घ ञ झ घ ध न श ल र ऊ च न द भ न ः त न ब घ ञ झ घ ध न श ल र ऊ च न द भ न ः त न ब घ ञ झ घ ध न श ल र ऊ च न द भ न ः त न ब घ ञ झ घ ध न श ल र ऊ च न द भ न ः त न ब घ ञ झ घ ध न श ल र ऊ च न द भ न ः त न ब घ ञ झ घ ध न श ल र ऊ च न
The merits of the proposed scheme - 1

- The encoding is Universal since it encompasses all characters that are found in general Tamil text interchange.

- The encoding is very efficient to parse.

For example

<table>
<thead>
<tr>
<th>Character</th>
<th>Code Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>ங</td>
<td>xx10</td>
</tr>
<tr>
<td>+</td>
<td>xx0C</td>
</tr>
<tr>
<td>கணக</td>
<td>xx1C</td>
</tr>
</tbody>
</table>
The merits of the proposed scheme - 2

- By simple arithmetic operation the characters can be parsed

\[ xx10 + xx0C = xx1C \]

\[ \text{க்} + \text{து} = \text{தெது} \]

\[ xx1C - xx10 = xx0C \]

\[ \text{தெது} - \text{க்} = \text{து} \]
The merits of the proposed scheme - 3

- Sorting and searching is very simple.
  - The Collation is sequential in accordance with the code value
- The encoding is unambiguous
  - Any given code point always represents the same character.
  - There is NO ambiguity as in the Present Unicode Tamil
Conclusion

- With the rapid spread of internet and search engines, Tamil language computing is at a critical stage.

- Government of Tamil Nadu and Government of India are seriously considering to set a standard that best meets the needs of Tamil in IT today and in the years to come.

- Visionary leadership at this stage will lead to wide spread use of Tamil in computers and is essential for the success of the e-governance efforts of the governments.

- This will also aid in enabling historians and archivists capture the public activities on the internet for posterity.

(Continued...)
Conclusion

- This computing revolution is very important. Failure to rectify the status of Tamil Unicode now will likely lead to the adoption of an incomprehensible encoding that might result in loss of information in the future.

- We should not regret in the future that what we store in the computers today will become unreadable.

- TACE-16 is the only alternative for efficient Tamil computing.

- The Tamil Nadu Government, therefore, strongly recommends to the Unicode Consortium for incorporating TACE-16 in the BMP space of Unicode.
Thank you