L2/07-175

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

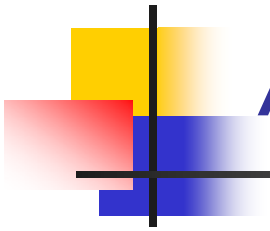
Mountain View, CA

Founding Exec.Committee Member, INFITT,

Member, Task Force on TACE-16

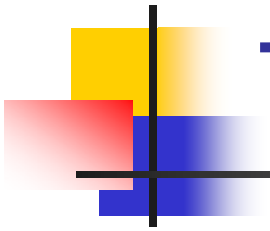
Chairman, Tamil Internet 2002 Conference, Foster City, CA.

Founder, TSCII.ORG.



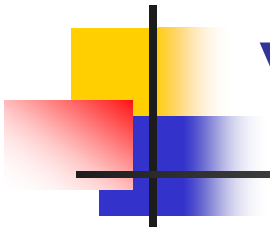
Agenda

- 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, true16-bit encoding
- TACE-16 Encoding and its merits
- Presentation, Testing and Reviews of TACE-16
- Proposal to Unicode



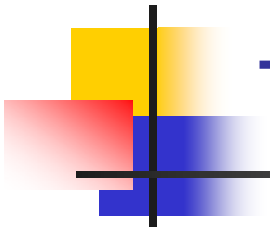
TACE-16 Task Force

- 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

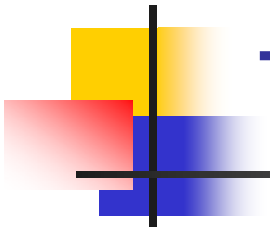


TN Government's Tamil IT initiatives - 1

- 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

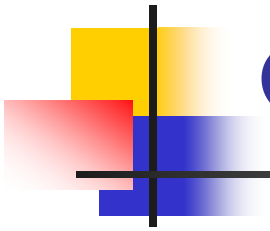
TN Government's Tamil IT initiatives - 2

- 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)

அ ஆ இ ஈ உ ஊ

எ ஏ ஐ ஒ ஓ ஔ

Aytham(1)

ஃ

Tamil Scripts

Consonants (18)

க்	ங்	ச்	ஞ்	ண்	ட்
த்	ந்	ப்	ம்	ய்	ர்
ல்	வ்	ழ்	ள்	ற்	ன்

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

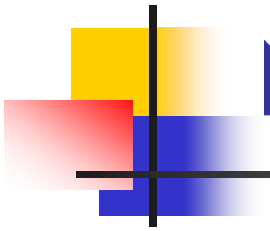
Uyir-Mey Characters *(Vowel Consonants)*

க்	க	கா	கி	கீ	கு	கூ	கெ	கே	கை	கொ	கோ	கௌ
ங்	ங	ஙா	ஙி	ஙீ	ஙு	ஙூ	ஙெ	ஙே	ஙை	ஙொ	ஙோ	ஙௌ
ச்	ச	சா	சி	சீ	சு	சூ	செ	சே	சை	சொ	சோ	சௌ
ஞ்	ஞ	ஞா	ஞி	ஞீ	ஞு	ஞூ	ஞெ	ஞே	ஞை	ஞொ	ஞோ	ஞௌ
ட்	ட	டா	டி	டீ	டு	டூ	டெ	டே	டை	டொ	டோ	டௌ
ண்	ண	ணா	ணி	ணீ	ணு	ணூ	ணெ	ணே	ணை	ணொ	ணோ	ணௌ
த்	த	தா	தி	தீ	து	தூ	தெ	தே	தை	தொ	தோ	தௌ
ந்	ந	நா	நி	நீ	நு	நூ	நெ	நே	நை	நொ	நோ	நௌ
ப்	ப	பா	பி	பீ	பு	பூ	பெ	பே	பை	பொ	போ	பௌ
ம்	ம	மா	மி	மீ	மு	மூ	மெ	மே	மை	மொ	மோ	மௌ
ய்	ய	யா	யி	யீ	யு	யூ	யெ	யே	யை	யொ	யோ	யௌ
ர்	ர	ரா	ரி	ரீ	ரு	ரூ	ரெ	ரே	ரை	ரொ	ரோ	ரௌ
ல்	ல	லா	லி	லீ	லு	லூ	லெ	லே	லை	லொ	லோ	லௌ
வ்	வ	வா	வி	வீ	வு	வூ	வெ	வே	வை	வொ	வோ	வௌ
ழ்	ழ	ழா	ழி	ழீ	ழு	ழூ	ழெ	ழே	ழை	ழொ	ழோ	ழௌ
ள்	ள	ளா	ளி	ளீ	ளு	ளூ	ளெ	ளே	ளை	ளொ	ளோ	ளௌ
ற்	ற	றா	றி	றீ	று	றூ	றெ	றே	றை	றொ	றோ	றௌ
ன்	ன	னா	னி	னீ	னு	னூ	னெ	னே	னை	னொ	னோ	னௌ

Slide 13

S2 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



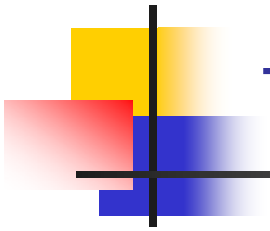
Nature of Tamil Vowel-Consonants

- 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, kI, etc., are characters to Tamils.
- This is the basis for the development of Tamil All Character Encoding scheme.

Grantha Letters

To represent Sanskrit borrowals

ஜ்	ஜ	ஜா	ஜி	ஜீ	ஜு	ஜூ	ஜெ	ஜே	ஜை	ஜொ	ஜோ	ஜௌ
ஶ்	ஶ	ஶா	ஶி	ஶீ	ஶு	ஶூ	ஶெ	ஶே	ஶை	ஶொ	ஶோ	ஶௌ
ஷ்	ஷ	ஷா	ஷி	ஷீ	ஷு	ஷூ	ஷெ	ஷே	ஷை	ஷொ	ஷோ	ஷௌ
ஸ்	ஸ	ஸா	ஸி	ஸீ	ஸு	ஸூ	ஸெ	ஸே	ஸை	ஸொ	ஸோ	ஸௌ
ஹ்	ஹ	ஹா	ஹி	ஹீ	ஹு	ஹூ	ஹெ	ஹே	ஹை	ஹொ	ஹோ	ஹௌ
க்ஷ்	க்ஷ	க்ஷா	க்ஷி	க்ஷீ	க்ஷு	க்ஷூ	க்ஷெ	க்ஷே	க்ஷை	க்ஷொ	க்ஷோ	க்ஷௌ
ஸ்ரீ												



Tamil Scripts

Total characters in Tamil including Grantha letters	: 325
Tamil Numerals	: 13
Special Characters	: 9
Total code points required	347

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

Tamil Scripts

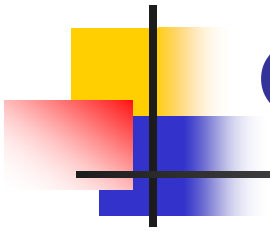
Usage of Tamil characters in plain text :

'இக்காலத் தமிழ்க் கவித இலக்கியத்தின் தந்த' என்று போற்றத் தகுந்தவர் பாரதியாரே என்பத நீங்கள் நன்கறிவீர்கள். காலத்தால் உருவாக்கப்பட்டுக் காலத்தப் புப்பிப்பவன் கவிஞன் என்று சொல்லப்படுவத அறிவீர்களா? இந்த இலக்கணத்க்கு முற்றிலும் பொருத்தமானவர் பாரதியார். பாரதிக்கு முந்திய காலம் எப்படி இருந்த? தமிழ் இலக்கிய வரலாற்றில் அ ஒரு தேக்கநில. பாரதிதாசன் குறிப்பிடுவ போலப் புலவர்கள், 'கலம்பகம் பார்த்தொரு கலம்பகத்தயும், அந்தாதி பார்த்தொரு அந்தாதி தன்னயும், மால பார்த்தொரு மால தன்னயும், காவியம் பார்த்தொரு காவியம் தன்னயும்' வரந் வந்தனர். புற்நீசல் போலத் தலபுராணங்கள் தோன்றின. இதில் தவறு என்ன என்று கேட்பீர்கள். இலக்கியம் முதன்மயாகப் படப்பாளியின் காலம், சூழல், அப்போதய சிந்தனப் போக்கு, படப்பாளியின் சொந் த உள்ளமப்பு, அடுத் வரும் காலம் பற்றிய அவன தொலநோக்கு ஆகியவற்ற எதிரொலிக்க வேண்டும் அல்லவா! மேலே குறிப்பிட்ட போல 'அரத்த மாவ அரக்கும்' புலவர்களிடம் இந்த இலக்கணத்த எதிர்பார்க்க முடியுமா? அந்தக் காலப் பகுதியில் குறிப்பிட்டுச் சொல்லும்படியான இலக்கியச் சாதனகள் தாயுமானவர், குமரகுருபரர், இராமலிங்கர் போன்றோரின் படப்புகளும் குறவஞ்சி, பள்ளு போன்ற இலக்கியங்களும் தாம். நாவல் படப்பு முயற்சியில் ஈடுபட்ட வேதநாயகம் பிள்ள, அ. மாதவயா, ராஜம் அய்யர் போன்றோரயும் சாதனப் பட்டியலில் சேர்த்க் கொள்ளலாம்.

திடுமெனக் கவிந்கொள்ளும் ஒரு புழுக்கம் ஒரு கனமழயக் கொண்டு வருவ போல், இலக்கியச் சிந்தனயிலும் வெளிப்பாட்டிலும் நேர்ந்த ஒரு புழுக்க நில பாரதி என்ற கனமழயக் கொண்டு வந்த. தமிழ் இலக்கியப் போக்கில் மிகப் பெரும் திருப்பங்கள உருவாக்கிய இருபதாம் நூற்றாண்டு. மேனாட்டார் தொடர்பு, அச்சு எந்திரம் போன்ற புறக் காரணங்களும், அரசியல் மற்றும் சமூக வாழ்வில் இந்தியர்களிடம் - தமிழர்களிடம் தோன்றிய விழிப்புணர்வு எனும் அகக்காரணமும் இருபதாம் நூற்றாண்டு தொடங்குமுன்பே இலக்கிய மறுமலர்ச்சிக்கான விதகளத் தூவிவிட்ட உண்மயே. எனினும் இருபதாம் நூற்றாண்டின் தொடக்கத்தில், சரியான இலக்குகள நோக்கிய தமிழ்ப் புத்திலக்கியம் பாரதியிடமிருந்தான் தொடக்கம் கொள்கிற. எப்படி என்று கேட்கிறீர்களா?

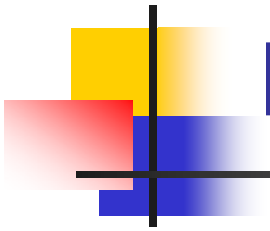
பாரதி 'சிந்க்குத் தந்த' மட்டும் அல்லர்; புதிய சிந்தனக்கும் தந்த அவர். தமிழில் அவர இல்லாத புதிய கருத்தோட்டங்கள வெள்ளமெனக் கொண்டு வந் சேர்த்தார். புலவர்களால் புறக்கணிக்கப்பட்டு வந்த பொமக்களின் மொழியக் கவிதயாக்கி அ எந்த அளவு பாய்ச்சல் தன்ம கொண்ட என்பத நிறுவிக் காட்டினார். மேலே நீங்கள் பார்த்தீர்களே, 'காலத்தப் புப்பிப்பவன் கவிஞன்' என்று, அதன அப்படியே செயல்படுத்திக் காட்டினார். படப்பாளி நேரடியாகச் சமூகத்த எதிர்கொள்பவன், தயக்கமில்லாமல் திறனாய்வு செய்பவன், அதன் குறகள, இழிவுகள அங்கதம் செய்பவன், சமூக மாற்றத்திற்கான சீர்திருத்தக் கருத்கள உரத்ச் சொல்பவன்-என்பன போன்ற இலக்கணங்களுக்கு முதல் பெரிய இலக்கியமாகத் திகழ்ந்தார். 'கடவிரித்தேன், கொள்வாரில, கட்டிக்கொண்டேன்' என்ற மனநில பாரதிய ஒருபோம் தொட்டதில்ல. 'நமக்குத் தொழில் கவித, நாட்டிற்குழத்தல், இமப்பொழும் சோராதிருத்தல்' என உறுதி கொண்டார். 'வயம் தழக்க வப்பேன். அமரயுகம் செய்யத் னிந் நிற்பேன்' எனத் தம் கவிதயின் ஆற்றலில் அசயாத நம்பிக்க கொண்டார். 'மண் பயனுற வேண்டும்', 'வானக மிங்குத் தென்பட வேண்டும்' என்ற பெருநோக்கம் கொண்டார்.

பாரதியாரின் கவிதகள் அவற்றின் உள்ளடக்க வகமயின் அடிப்படயில் தோத்திரப் பாடல்கள், வேதாந்தப் பாடல்கள், தேசிய கீதங்கள், காவியங்கள், பல்வகப் பாடல்கள் எனப் பல பிரிவுகளாக வெளியிடப்பட்டுள்ளன. ஒவ்வொரு வகயிலிருந்ம் தெரிந்தெடுத்த பாடல்கள் உங்கள் பாடப்பகுதியில் சேர்க்கப்பட்டுள்ளன. பாரதிய முழுமயாகப் பார்க்க இப்பாடம் உங்களுக்கு ணபுரியும். முதலில் பாரதியின் வாழ்வும் படப்பும் பற்றிப் பார்ப்போம்.



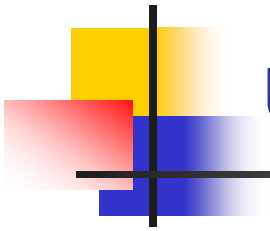
Current Tamil Encodings

- ISCII – 7 bit
- TSCII/TAB – 7bit
- 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

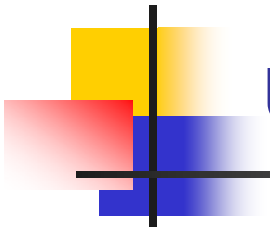


Unicode Design Goals

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.

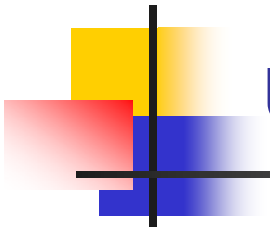


Unicode Design Goals

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

	0B8	0B9	0BA	0BB	0BC	0BD	0BE	0BF
0		ஐ		ர	ீ			ய
1				ற	ு			ர
2	்	ஒ		ல	ூ			ற
3	ஃ	ஓ	ண	ள				
4		ஔ	த	ழ				
5	அ	க		வ				
6	ஆ				ெ			
7	இ			ஷ	ே	ௗ	க	
8	ஈ		ந	ஸ	ை		உ	
9	உ	ங	ன	ஹ			ந	
A	ஊ	ச	ப		ொ		ச	
B					ோ		ரு	
C		ஜ			ௌ		சு	
D					◌		எ	
E	எ	ஞ	ம	ா			அ	
F	ஏ	ட	ய	ி			கூ	

- 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

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.

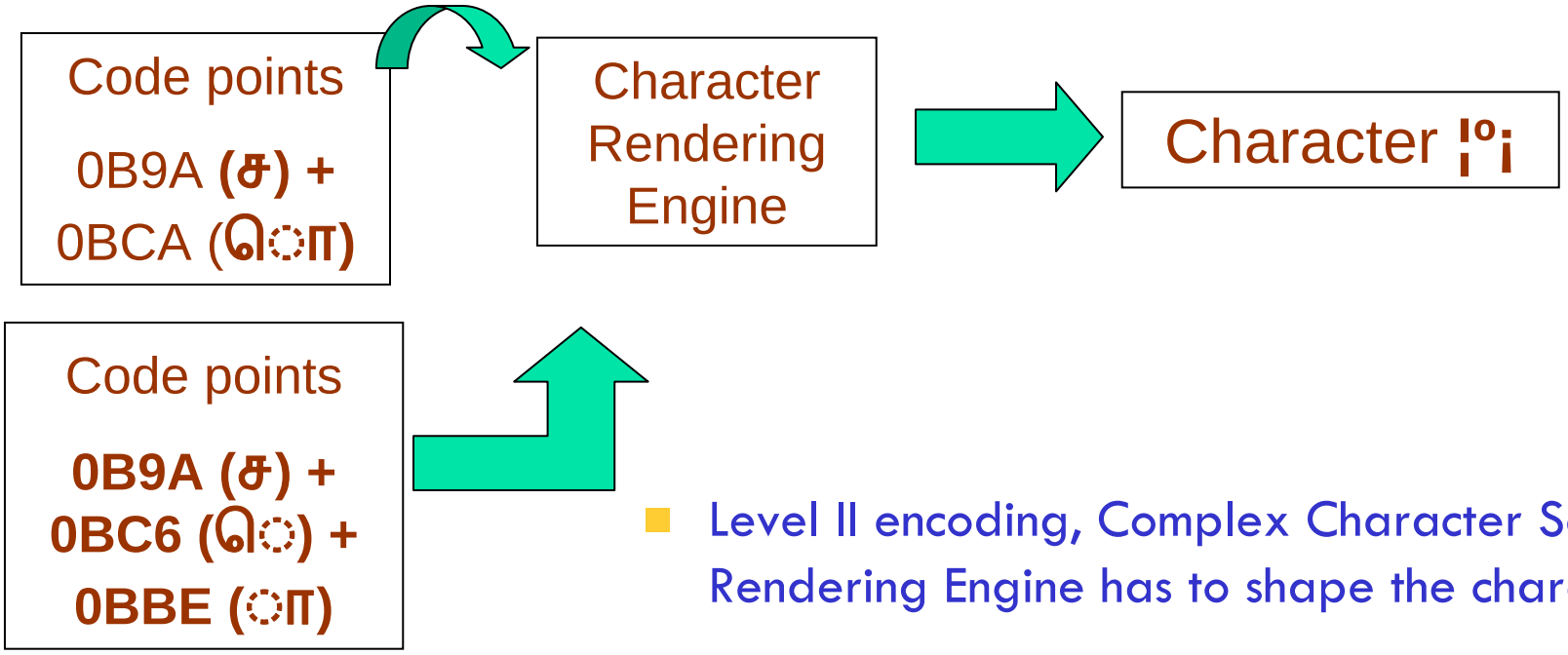
கை, ஙை, ..., னை

Violation of Unicode principles in the Present Unicode Tamil Encoding

- There are two methods of rendering the following Vowel Consonants

This leads to ambiguity in rendering characters

Rendering of Vowel Consonants



- 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:

  ம ண ி வ ண ் ண ன ்

- 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

Investigation Type	SCHEME 1 (Unicode 3.0)	SCHEME 2 (Consonant-Vowel)	Scheme 3 (All Character)
1. Data storage, retrieval and display parameters			
File Size	147	142	100
Display time	2,500	2,875	100
File transfer time	147	142	100
Find & replace time	270	257	100
2. Database related parameters			
DB size	120	118	100
DB creation time	112	112	100
Indexed DB size	142	141	100
DB indexing time	178	160	100
DB sorting time	164	147	100
DB record search	103	108	100
3. Morphological analysis parameters			
Morphological analysis	526	284	100
Noun search time	476	357	100
Verb search time	208	150	100
Gender search time(1)	185	172	100
Gender search time(2)	158	152	100

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

- பர்க்கெலி தமிழ்ப் பேராசிரியர் ◆fார்ˆ +ார்ட்,
- பென்சில்வேனியா பேரா. "ிப்மன்
- ஒரு காலத்தில் ராfாவெல்லாம் முகமூடி போட்டு தனது ராfாங்கத்தின் நிர்வாகத்தை தானே சென்று பார்த்து தவறுகளை உடனுக்குடனேயே செய்து வந்தனர்
- தமிழர்கள் ஆ?ா ஓ◆?ா என்று பேசுவார்கள்

More Tamil Unicode examples

"வீட்டின் முழு உரிமையையும் திருமதி ரோஜாவுக்குக் கொடுத்து விடுகிறேன்"

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

"வீட்டின் முழு உரிமையையும் திருமதி ரோ◌ாவுக்குக் கொடுத்து விடுகிறேன்"

Will the court give the property to

திருமதி ரோஜா or திருமதி ரோஷா?

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 ங .....

- The following Tamil scripts

கா ஙா ..... னா

:

:

கொ ஙொ ..... னொ

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. க் + ஆ = கா

க் + ஒ = கொ

Unicode provides for rendering கொ க் + ெ + ா = கொ

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)

16-பிட்டு தமிழ் அனைத்துரு குறியீட்டு முறை

Annexure - 1

	xx0	xx1	xx2	xx3	xx4	xx5	xx6	xx7	xx8	xx9	xxA	xxB	xxC	xxD	xxE	xxF	xy0	xy1	xy2	xy3	xy4	xy5	xy6	xy7	xy8	xy9	xyA	xyB
0		க்	ங்	ச்	ஞ்	ட்	ண்	த்	ந்	ப்	ம்	ய்	ர்	ல்	வ்	ழ்	ள்	ற்	ன்	ஜ்	ஸ்	ஷ்	ஸ்	ஹ்	க்ஷ்		ం	ఀ
1	அ	க	ங	ச	ஞ	ட	ண	த	ந	ப	ம	ய	ர	ல	வ	ழ	ள	ற	ன	ஜ	ஸ	ஷ	ஸ	ஹ	க்ஷ		క	మీ
2	ஆ	கா	ஙா	சா	ஞா	டா	ணா	தா	நா	பா	மா	யா	ரா	லா	வா	ழா	ளா	றா	னா	ஜா	ஸா	ஷா	ஸா	ஹா	க்ஷா		ఖ	ఁ
3	இ	கி	ஙி	சி	ஞி	டி	ணி	தி	நி	பி	மி	யி	ரி	லி	வி	ழி	ளி	றி	னி	ஜி	ஸி	ஷி	ஸி	ஹி	க்ஷி		గ	ం
4	ஈ	கீ	ஙீ	சீ	ஞீ	டீ	ணீ	தீ	நீ	பீ	மீ	யீ	ரீ	லீ	வீ	ழீ	ளீ	றீ	னீ	ஜீ	ஸீ	ஷீ	ஸீ	ஹீ	க்ஷீ		ఘ	ః
5	உ	கு	ஙு	சு	ஞு	டு	ணு	து	நு	பு	மு	யு	ரு	லு	வு	ழு	ளு	று	னு	ஜு	ஸு	ஷு	ஸு	ஹு	க்ஷு		ఙ	ఄ
6	ஊ	கூ	ஙூ	சூ	ஞூ	டூ	ணூ	தூ	நூ	பூ	மூ	யூ	ரூ	லூ	வூ	ழூ	ளூ	றூ	னூ	ஜூ	ஸூ	ஷூ	ஸூ	ஹூ	க்ஷூ		చ	అ
7	எ	கெ	ஙெ	செ	ஞெ	டெ	ணெ	தெ	நெ	பெ	மெ	யெ	ரெ	லெ	வெ	ழெ	ளெ	றெ	னெ	ஜெ	ஸெ	ஷெ	ஸெ	ஹெ	க்ஷெ		ఎ	ఆ
8	ஏ	கே	ஙே	சே	ஞே	டே	ணே	தே	நே	பே	மே	யே	ரே	லே	வே	ழே	ளே	றே	னே	ஜே	ஸே	ஷே	ஸே	ஹே	க்ஷே		ఞ	
9	ஐ	கை	ஙை	சை	ஞை	டை	ணை	தை	நை	பை	மை	யை	ரை	லை	வை	ழை	ளை	றை	னை	ஜை	ஸை	ஷை	ஸை	ஹை	க்ஷை		ఫ	
A	ஒ	கொ	ஙொ	சொ	ஞொ	டொ	ணொ	தொ	நொ	பொ	மொ	யொ	ரொ	லொ	வொ	ழொ	ளொ	றொ	னொ	ஜொ	ஸொ	ஷொ	ஸொ	ஹொ	க்ஷொ		ఠ	
B	ஓ	கோ	ஙோ	சோ	ஞோ	டோ	ணோ	தோ	நோ	போ	மோ	யோ	ரோ	லோ	வோ	ழோ	ளோ	றோ	னோ	ஜோ	ஸோ	ஷோ	ஸோ	ஹோ	க்ஷோ		ఱ	
C	ஔ	கௌ	ஙௌ	சௌ	ஞௌ	டௌ	ணௌ	தௌ	நௌ	பௌ	மௌ	யௌ	ரௌ	லௌ	வௌ	ழௌ	ளௌ	றௌ	னௌ	ஜௌ	ஸௌ	ஷௌ	ஸௌ	ஹௌ	க்ஷௌ		ఴ	
D	ஃ																								ஶ்ரீ			
E																												
F																												

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

Character	Code Value
க்	xx10
+	
ஔ	xx0C
↓	
கௌ	xx1C

The merits of the proposed scheme - 2

- By simple arithmetic operation the characters can be parsed

xx10 + xx0C = xx1C

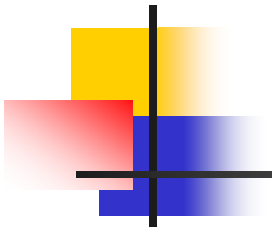
க் + ஔ = கௌ

xx1C - xx10 = xx0C

கௌ - க் = ஔ

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

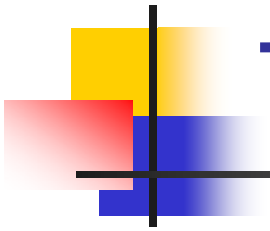
- With the rapid spread of internet and search engines, Tamil language computing is at a critical stage.
- Government of Tamil Nadu and Government of India are seriously considering to set a standard that best meets the needs of Tamil in IT today and in the years to come.
- Visionary leadership at this 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

(Continuation)

- 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

Thank you