

REPORT ON THE FINAL RECOMMENDATIONS OF THE TASK FORCE ON TACE16

1. PREAMBLE

Initiatives for the usage of Tamil in computers and in Information technology started as early as 1985. The Tamils living in different parts of the world started developing their own 8-bit bi-lingual encodings for Tamil using the extended ASCII area. Thus, there were a number of encodings for Tamil in use around the world. Mean time the Government of India introduced a 7-bit encoding called ISCII for all the Indian languages for transliteration purposes. In October 1991 Unicode Tamil was announced by the Unicode Consortium incorporating the 7-bit encoding of ISCII standard in the 16-bit environment of Unicode. But this standard of Unicode Tamil was not put to use until recently. The usage of many encoding schemes for Tamil posed a number of problems for the users as well as developers, causing a big concern for the Tamil computing world. This problem was discussed in an International conference **TamilNet'97** held in Singapore in May 1997. It was resolved in the conference that the Keyboard and Encoding diversities should be solved and the Tamil Nadu Government should take initiatives to evolve 8-bit Encoding and Keyboard standards for Tamil.

Accordingly, the Tamil Nadu Government constituted a Sub-Committee on Tamil in Information Technology with Dr.M.Anandkrishnan as its Chairman, through G.O (Ms) No.653 dated 08.10.1998 under the State Task Force for Information Technology headed by the Chief Minister, for solving the above problem. The Task Force organized an International Conference and seminar on Tamil in Information Technology - **TamilNet'99** during 7th and 8th February 1999. During the conference through discussions and deliberations a phonetic Keyboard Standard was evolved and a Bi-lingual-TAB and a Mono-lingual-TAM encoding standards for Tamil were recommended for evaluation. After an evaluation period of 100 days these standards were declared by the Tamil Nadu Government in a G.O.(Ms) No. 17 dated 13.06.1999. Like TAB encoding Unicode Tamil was also an 8-bit encoding incorporated in 16-bit environment, limiting the number of Tamil Characters encoded to only 36 characters, while 16-bit encoding was meant for providing code space for all characters of a language used in text. The Unicode Tamil also had many other problems. During the **TamilNet'99** conference the issues related to the Unicode Tamil were also discussed. A detailed account of the deficiencies of the Unicode Tamil is given in **Annexure-1**. Based on the recommendations made in the conference, the Government of Tamil Nadu, in the above G.O., has directed the Sub-Committee on Tamil in Information Technology to propose an efficient 16-bit

character encoding scheme for adoption in the Unicode. The G.O. further says that Tamil Nadu Government have become an Associate Member of the Unicode Consortium, USA in order to facilitate the submission of a revised character encoding standard for Tamil to the Unicode Consortium.

The Sub-Committee on Tamil in Information Technology entrusted this responsibility to the Tamil Virtual University (TVU) to evolve a suitable Character encoding scheme for Tamil for adoption into the Unicode standard through appropriate testing and evaluation of the possible schemes, using the Tamil Software Development Fund.

TVU formed a committee with experts pooled from KaNithamizh Sangam for evolving a suitable Character encoding scheme for Tamil for adoption into the Unicode standard. The committee developed a 16-bit All Character encoding for Tamil and the same was presented at the pre-conference session of TamilNet 2000 conference in Colombo, Sri Lanka as well as at the main TamilNet2000 conference in Singapore. This was also discussed at the TamilNet 2001 conference in Malaysia where an expert from Microsoft was present. The problem of Unicode Tamil was also discussed widely in a work group of INFITT. Subsequently the new scheme was presented at a meeting convened by MIT, GOI, on 2nd November 2000 and the same was submitted in the prescribed format to the Ministry of Information Technology, for onward submission to the Unicode consortium. Dr.Om Vikas, the then Director, MIT, presented the same in the Technical Committee meeting of the Unicode Consortium, held during 7-10 November 2000 at Sandiago, USA. The Unicode consortium deliberated on the same and observed in their UTC document L2/01-430 that the proposed scheme should be justified by the results of scientific studies for consideration to include in the Unicode standard.

Accordingly, an all Character 16-bit encoding scheme for Tamil, and a vowel and consonant scheme along with the present Unicode Tamil were tested and evaluated on different applications. The results of the investigation were favorable to the 16-bit All Character Encoding for Tamil (**Annexure-2**).

The MIT GOI had discussions on the new scheme in its meetings held in September 2001 and in November 2001 and submitted the same to the Unicode Consortium for discussion in the Unicode Technical Committee (UTC), held in USA during November, 2001. The UTC, in its meetings, argued that UTC would like to work with the experts at MIT and INFITT to show that the current Unicode Tamil encoding can represent all Tamil syllables, and that the Unicode Collation Algorithm can be used, with the appropriate tailoring, to correctly order Tamil words. If other encodings of Tamil are developed in the future, the UTC would work together with the appropriate organizations to develop precise mapping tables between those encodings and Unicode. Unfortunately, representatives from Tamil Nadu Government could not participate in such the UTC meetings to present the

issues and requirements of Unicode encoding for Tamil effectively in the UTC meetings.

On 16th July 2002 MIT called for a meeting of the Language and computing experts to debate on this specific issues of providing 24 x 16 code points for the representing Tamil language in the Unicode. The meeting was attended by Dr.Ponnaivaikko, Director, Tamil Virtual University, Mr. Hari an expert from IBM, Dr. M.N. Cooper, Joint Director of Modular InfoTech and Mr.N. Anbarasan, Managing Director of AppleSoft along with MIT personnel and language experts. After a detailed discussion, the expert members uniformly agreed that the 16 bit Unicode proposed for Tamil is an excellent scheme and they further recommended that similar schemes should be adopted for other Indian Languages also. Based on the recommendations of the experts, it was resolved to assign the job of evolving a similar 16 bit encoding for all Indian Languages to the Consortium for innovation in language technology (CoIL).

To obtain the views of the Tamil Diaspora on the All Character Encoding for Tamil the test results were uploaded in the website www.tunerfc.tn.nic.in requesting for comments from the Tamil Diaspora. The new encoding was named as TUNE (Tamil Unicode New Encoding). The proposed all character encoding scheme was placed in the Private Use Area (PUA) of Unicode in the Unicode block **E200 to E38F** on 24th June 2005 so as to put the scheme in use by the Tamil Diaspora. For the purpose of testing and evaluating the new encoding in PUA under various operating platforms and in all possible applications, a new keyboard driver and a new Unicode font for TUNE were developed. The feedback from the Tamil Diaspora was very much encouraging and they had reported that the TUNE is as efficient as English in all applications and it is at least 40% to 200% more efficient than the current Unicode standard version 4.0. A draft report on TUNE RFC is available at www.tunerfc.tn.nic.in

Dr. M. Ponnaivaikko visited US and Canada during June/July 2006 and participated in the FETNA conference held during 1st to 3rd July 2006 in New York and addressed the Tamil Community and also had discussions with the Software professionals in the US and Canada. The Tamil software professionals in US and Canada fully support TUNE. FETNA organizing Committee passed resolutions to the effect that “the delegates to the Federation of Tamil Associations of North America Convention meeting in New York, July 3, 2006, urge the Union Government of India and the State Government of Tamil Nadu, to recognize the TUNE encoding as the standard 16-bit encoding for Tamil Language; that the Union Government of India and the State Government of Tamil Nadu be urged to enforce the TUNE encoding as the Indian national standard for Tamil 16-bit encoding in a way that will restore trust among the Tamil speakers; that the Union government of India be urged to enforce the TUNE encoding in all the Tamil software that is sold to the Central and State Governments of India; and that the

Tamil language users on computers and internet be urged to follow the TUNE encoding as a standard in their use of Tamil in computers” (Copy of the resolution is given in **Annexure-3**). A conference on Tamil 16-bit All Character Encoding was organized on 2nd Sept. 2006 by TVU to consolidate the views and comments received on the Tamil 16-bit All Character Encoding placed in the PUA of the Unicode space and to plan further course of action to declare one single true 16-bit all character encoding scheme for Tamil as a standard in the place of existing 8-bit encoding standards and for moving the same into Unicode. The conference was inaugurated by the Honorable Minister for Communication and Information Technology, Government of India, The conference was attended by delegates from Singapore and SriLanka. Software professionals from major firms such as IBM, MICROSOFT and Tamil software developers in Tamil Nadu and in the other parts of the country and abroad participated. In his inaugural address the Honorable Minister said that the new scheme shall be reviewed and revised based on the comments received and tested on different platforms and in different applications like E-Governance, web publication, Natural Language Processing, etc. The Minister further said that creation of a corpus fund will be considered for testing and development and for encouraging migration and conversion to the new encoding. Minister desired that the 16-bit encoding for Tamil shall be made available soon for implementation in the e-Governance project by the Government.

The outcome of the deliberations in the conference led to an unambiguous and unique consensus that there should be only one encoding for Tamil and that should be the 16-bit Tamil all character encoding. Consensus was also arrived at for evolving an implementation strategy to achieve this goal within a specified time frame. In The conference made the following Recommendations for the consideration of the Government of Tamil Nadu:

- i. The Government of Tamil Nadu may consider formation of a Task Force to coordinate the activities related to the development of an acceptable 16-bit All Character Encoding for Tamil Language, through appropriate testing and validation with the following mandates.
- ii. The Government of Tamil Nadu may take necessary action to publicize the proposed the 16-bit Tamil All Character scheme of encoding in the countries where Tamil is an official language so as to get their comments on the proposed scheme.
- iii. The Government may create a corpus fund for providing financial and policy support for migrating contents and developments already done in the current environment.

- iv. The corpus fund created may include funds for developing tools and drivers to support the 16-bit Tamil All Character encoding in different platforms such as Windows, Macintosh, Linux, and UNIX for free distribution.
- v. The Government of Tamil Nadu may become a Full voting Member of Unicode Consortium so that the State can directly submit proposals to Unicode consortium for adopting Tamil-16 bit All Character encoding into Unicode.

Based on the recommendations of the conference held for consolidation, the Government of Tamil Nadu constituted a Task Force in the G.O.(Ms)No.13 Information Technology Department dated 10.11.2006 (**Annexure-4**), under the Chairmanship of Dr.M.Anandakrishnan with Dr.M.Ponnaivaikko as the Vice-Chairman, Dr.P.R.Nakkeeran as the convener and 10 other experts as Members, to formulate action plan for the implementation of the following recommendations made in the conference held on 2nd September 2006:

- ❖ Action to publicize the proposed Tamil-16 bit All Character scheme of encoding in the countries where Tamil is an official language so as to get their comments on the proposed scheme.
- ❖ To create a corpus fund for providing financial and policy support for migrating contents and developments already done in the current environment.
- ❖ The corpus fund to be created shall include funds for developing tools and drivers to support Tamil-16 bit All Character encoding in different platforms such as windows, Macintosh, Linux, and UNIX for free distribution.
- ❖ The Government of Tamil Nadu shall become a Full voting Member of Unicode Consortium so that the State can directly submit proposals to Unicode consortium for adopting Tamil-16 bit All Character encoding into Unicode.

The Committee took the job of evaluating the 16-bit All Character Tamil encoding by assigning the job to the competent testing agencies and deliberated on the test results and on other issues in its several meetings. The findings and the recommendations of the Committee are presented in this Report.

2. ACTION TAKEN BY THE TASK FORCE

The Task Force in its first meeting decided the following course of action:

- The 16 bit all character-encoding scheme (TACE16 in **Table-1**) which is already available in the Private User area can be tested without any change for certain applications.
- There is a need to consider the existing Unicode Tamil scheme also for testing and comparing the results.
- There is a need to test thoroughly all the schemes before coming to any conclusion.
- The test areas should be in the applications of
 - (i) e-Governance - with internet and intranet
 - (ii) Natural Language Processing.
 - (iii) Publishing.
- Preparing the action plan, conducting the tests and monitoring the progress will be supervised by one of the members of the Task Force for each testing area as identified bellow :
 - (i) e-governance, browsers - Thiru. A.Mohan, NIC
 - (ii) publishing - Dr.M.N.Cooper,
Modular Infotech.
 - (iii) Natural Language Processing - Prof.V.Krishnamoorthy,
Crescent Engineering College.
- Mr.N.Anbarasan will present possible modifications for the existing scheme for improvements, if any.
- Transparencies in testing should be emphasized in all aspects.

The Task Force met 13 times during the period from December 2006 to January 2008 to discuss the results of the tests carried out by the different investigators.

2.1. Presentation of TACE16 in the UTC Meeting

In May 2007 the Government of Tamil Nadu became a voting member of the Unicode Consortium and submitted a proposal for adopting TACE16 in the

Unicode. Dr.M.Ponnaivaikko and Thiru.Mani M.Manivannan along with Pankaj Agrawala, Joint Secretary, MICT, Gov't of India, participated in the UTC meeting held during 14th to 18th May 2007 and presented the proposal. After detailed deliberations in many sittings during the meeting from 14th to 18th May, it was decided to set up a subcommittee to examine the encoding issues of Tamil and other scripts of India with Mr.Eric Muller, UTC Vice Chair -- Adobe, San Jose, CA as the Chair of the Subcommittee (**Annexure-5**) with following goals:

- i. Study, review, and document the current Tamil Unicode Representations.
- ii. Identify the stability issues with respect to TACE16
- iii. Identify solutions to bridge limitations of (1) with the advantages of TACE16.
- iv. Identify ways to accommodate TACE16 in BMP.
- v. Identify ways to interoperate with TACE16 (interoperable standard).

The following decisions were made in respect of the functioning of the Subcommittee:

- i. The subcommittee will discuss through an e-group mail on the issues.
- ii. Subcommittee will have teleconference meetings every month.
- iii. The subcommittee will set up a mailing list to discuss Tamil and other Indic languages.
- iv. The subcommittee will promote exchange of documents through a web page.
- v. The subcommittee will encourage member organizations to nominate their participants in the mailing list and sponsor their experts.
- vi. The scope of the work will not be limited to character encoding; it will address general international issues implementation issues, CLDR etc.
- vii. The UTC will continue to discuss Tamil in the upcoming Meetings.
- viii. The subcommittee will meet in Chennai in December 2007 to deliberate the findings and decisions of the subcommittee.

During the UTC meeting it was pointed out that there are 484 free spaces in the BMP area of the Unicode space which can be used for accommodating the Tamil

characters. Making use of these free locations, the characters of TACE16 were assigned new locations accommodating all the Tamil characters in 6 blocks as in **Table-2**.

The testing teams were then requested to study this scheme also as New TACE16 along with the earlier TACE16 designating as old TACE16.

The study reports of the three testing agencies are appended to this Report. A brief discussion of the outcome of the studies is given below.

2.1.1. Test Results on E-Governance and Browsing:

Old and New TACE16 and the present Unicode Tamil were tested for Data storage, Sorting, Searching and online Data entry into a web page. The following are the observations:

- TACE16 is efficient over Unicode Tamil by about 5.46 to 11.94 percent in the case of Data Storage Application.
- TACE16 is efficient over Unicode Tamil by about 18.69 to 22.99 percent in the case of Sorting Index Data.
- TACE16 is efficient over Unicode Tamil by about 25.39% when the entire data is of Tamil.
- The default collation sequence followed (Binary) while using the code space values in the New TACE16 is not as per Tamil Dictionary order. Some of the uyir-meys (Agara-uyirmeys) are taking precedence over vowels and other Uyir-meys in the New TACE16, the vowels and agarauyir-meys being in the 0B80 - 0B8F block and the other Uyir-meys being in the 0800 to 08FF. Because of this reason, sorting Unicode data looks better than TACE16 data.
- TACE16 is faster in sorting over Unicode by about 0.31 to 16.96 percent.
- Index creation on TACE16 data is faster by 36.7% than Unicode.
- For Full key Search on Indexed Fields, TACE16 performed better than Unicode by upto 24.07%. In the case of non-indexed fields also TACE16 performed better than Unicode by upto 20.9%.
- Rendering of static Tamil Data was fine with TACE16.

- It was not possible to do data entry satisfactorily with New TACE16 using online data entry forms in the web applications, primarily due to the fact that some of the characters are placed in Arabic code area. However, data entry using Old TACE16 was seen to be proper in both IE and FireFox Browsers.

2.1.2. Test Results on Publication Applications

Old and New TACE16 and the present Unicode Tamil were tested in various applications for different criteria like installation, un-installation of fonts, full character set typing, hyphenation, find-and-replace function, indexing function, printing on different printers etc. Three main platforms were chosen, Windows, Linux and Mac. Applications under each platform were identified.

The following are the observations based on the results of the tests carried out for Old TACE16:

- In all the applications tested on various platforms, only Microsoft office and Open office are enabled with Tamil Unicode. All other applications could not work with Unicode font. However, in almost all applications, TACE16 font and Keyboard handler worked properly.
- It was not possible to type in Tamil with TACE16 font in all of the dialog boxes like Find/Replace, Spell-Checker etc. To overcome this problem, OS manufacturers and application suppliers should enable the relevant dialog boxes to accept TACE16 font. Or all language related support will have to be provided by adding plugins to the specific applications, like find/replace, spell-checker, hyphenation etc,
- TACE16 font and keyboard handler does not work in Windows XP and Microsoft PowerPoint. Microsoft can make it possible if TACE16 is accepted by the Unicode Consortium.
- The raw text entered in TACE16 encoding requires about 30% less memory than that of the same text material entered in Tamil Unicode. Functions such as Find/Replace (find and replace text strings were pasted in the dialog boxes), opening/closing of applications etc were quite fast in case of TACE16.

The following are the observations based on the results of the tests carried out for New TACE16:

The test cases that were used in testing Old TACE16 were used for testing the New TACE16 also.

- In the very early phase of testing on Windows (XP, 2003 and Vista) a very peculiar behavior was noticed. This was specially noticed in applications that were enabled with Tamil Unicode (Microsoft Office and Open Office). While data entry is made it took consonants and vowels from the Windows default Tamil font 'Latha' where as the remaining CV combinations were rendered correctly from TACE16 fonts. This behavior was attributed to interference of the existing language software with the TACE16. It was therefore decided to Disable Complex Script module in Windows XP and Uninstall 'Latha' font in Windows 2003 and Windows Vista. This behavior was not observed in Mac or Linux environment.
- Applications like Adobe Illustrator CS2 and Photoshop CS2 did not respond to code blocks UxAA60 to UxAA7F, UxAAE0 to UxAAFF, UxABE0 to UxABFF.
- In OpenOffice Writer applications, the typing proceeded from right-to-left direction for many characters.
- The raw text entered in New TACE16 encoding requires about 30% less memory than that of the same text material entered in Tamil Unicode. Functions such as Find/Replace (find and replace text strings were pasted in the dialog boxes), opening/closing of applications etc were quite fast in case of TACE16.
- It was not possible to type in Tamil with New TACE16 font in all of the dialog boxes like Find/Replace, Spell-Checker etc. as in the case of Old TACE16.
- If TACE16 is given a continuous code space in BMP area the TACE16 will out perform the current Tamil Unicode. All Character encoding has an inherent advantage over Unicode that it can enable any application at the developer end.
- The odd behavior observed during testing phase, may not be attributed to encoding but to the shortfalls in implementation of applications. This will vanish once the code block/s is/are regularized.
- The 30% compaction in the TACE16 encoding does not show up in file sizes saved by various applications due to compression algorithms used by these applications.

2.1.3. Test Results on NLP Applications

The aim of this NLP testing is to evaluate the efficiency of different character encoding schemes of Tamil using the Language technology test bed. The following two basic applications in language technology are chosen to evaluate Old and New TACE16 and the present Unicode Tamil encodings:

- i. Morphological Analyzer (MA)
- ii. Morph Generator (MG).

The following parameters were considered while evaluating the time taken by different encoding schemes:

- i. Test File Size - in number of words, in memory size
- ii. Dependent Dictionaries Sizes
- iii. Word Type - Nouns, Verbs, Adverb, Adjective
- iv. Word Length - 5 length to 23 Length words
- v. Type of NLP application - Morph Analyzer, Morph Generator
- vi. Type of OS - Windows XP, Windows Vista, Windows 2000, Linux

The evaluation is done on Linux and Windows operating systems. Sample data from different domains and in the three different encoding schemes were used. The following experiments were performed to test the efficiency of the encoding schemes.

- First experiment : Single word analysis in Unicode, Old TACE16 and New TACE16.
- Second experiment : 10000 words in Unicode, Old TACE16 and New TACE16.
- Third experiment : 25000 words in Unicode, Old TACE16 and New TACE16.
- Fourth experiment : 50000 words in Unicode and New TACE16 formats.

The Report on the test results is appended to this report. The following are the observations:

- ❖ Unicode takes 4 to 5 times more time than Old TACE and New TACE for the Initialization process in the case Single words.

- ❖ When the word level Analysis carried out, a text file is read, the ranges of the characters are verified to check whether they are Tamil characters or not. Since in the New TACE16, the characters are allocated in five different blocks, it has to be checked 4 times more than what is done in the case of old TACE16. When the input file size becomes larger, the computations also increase proportionately while reading. This increases the time complexity.
- ❖ New TACE16 and Old TACE16 take approximately the same time for analyzing the words in the experiment with word length. Unicode takes more time than the other two encodings.
- ❖ All the three encodings perform with less time in Windows compared to Linux.
- ❖ The above inferences are the same for both Morphological Analyser and Morphological Generator.

3. VISIT OF THE EXPERT TEAM FROM UNICODE CONSORTIUM

A team comprising Dr. Mark Davis and Michael Kaplan representing South Asia Subcommittee of the Unicode consortium visited Chennai to have discussions with the concerned officials of the Tamil Nadu Government, TVU and the members of the TASK FORCE on TACE16 on the issues related to the present Unicode Tamil. The Meeting took place on 23rd January 2008 at the Hotel Le Meridian, Guindy, Chennai and 24th January 2008 at Board Room of TVU. The meeting was attended by the Secretary, Department of IT, Government of Tamil Nadu, Joint Secretary, MIT, GOI, representative from INFITT, the members of the testing Agencies and the Chairman and Vice-Chairman of the Task Force along with the representatives from the Unicode Consortium. Various issues concerning the implementation TACE16 in the Unicode space were deliberated. It was insisted that TACE16 should be accommodated either in the BMP space or in the SMP space of the Unicode chart. Though the representatives were not in a position to make definite commitments, the assured the interest of the Government of Tamil Nadu will be reported to the Unicode consortium for a possible solution. The minutes of the meeting as a report submitted by Dr. Mark Davis is given in **Annexure-6**.

4. FEASIBILITY ASPECTS

Recognized as one of the Classical Languages of the World, Tamil is a rich language having at least 2500 years of Inscriptional records and literatures. Tamil is a Conservative Language and it preserves its continuity for millenniums of years. It has Alpha syllabic writing system including Vowels, Consonants and Vowel-Consonants, all with graphical representation as *single letters*. In the Unicode space Tamil language is not encoded in the right way preserving its true properties for efficient and effective use of the language in computers and information technology

as brought out in Section 1 of this report. Realizing this situation the Tamil Nadu Government initiated action to bring out an All Character 16-bit Encoding scheme for Tamil (TACE16) as a National Standard and for adoption into the Unicode Standard. This scheme has been tested and evaluated for various applications. The test results have indicated that this scheme would be the best, if incorporated into the Unicode space as a Standard. Difficulties were faced while testing this scheme, because of the limitations and constraints built into the application software systems of the MNCs, like Microsoft, Adobe, etc. Having convinced about the merits of the scheme TACE16, this section discusses about the feasibility of the same for implementation.

4.1. Feasibility as a National Standard

The Tamil Nadu Government declared in 1999 a Bi-lingual Standard TAB and a Mono-lingual Standard TAM for Tamil Language, the only Indian Language to have an 8-bit encoding standard. Over the past 8 years from 1999 till date, a number of Tamil software vendors have developed varieties of application software systems, including Tamil fonts, word processors, search engines, word nets OCRs, digital contents, dictionaries, system level tools and drivers, etc, using these encoding standards. These vendors may need to be supported with public domain conversion tools to help convert their software systems to the proposed encoding, if TACE16 is declared as a standard in place of TAB and TAM. The Tamil Diaspora is looking for one unique standard encoding for Tamil. The MNCs are keenly watching the developments in Tamil Nadu in respect of the 16-bit encoding standard for Tamil. The MNCs will not hesitate to changeover to this new Encoding Standard, if enough business opportunities are built for them in Tamil Nadu and in the Tamil Diaspora. The implementation of E-Governance in Tamil in the District and State administration and the use of Tamil in the administration of the departments, organizations, institution and the Universities in Tamil Nadu will be the motivating factors for the MNCs to implement TACE16 in their software systems. Seriousness will be felt in the issue only when TACE16 is declared as a National Standard. Technically there are no other problems for declaring TACE16 as a State and a National Standard.

4.2 Feasibility of TACE16 as a Unicode Standard

Two versions of TACE16 have been investigated, old and new versions.

4.2.1 Feasibility of Implementing the old TACE16

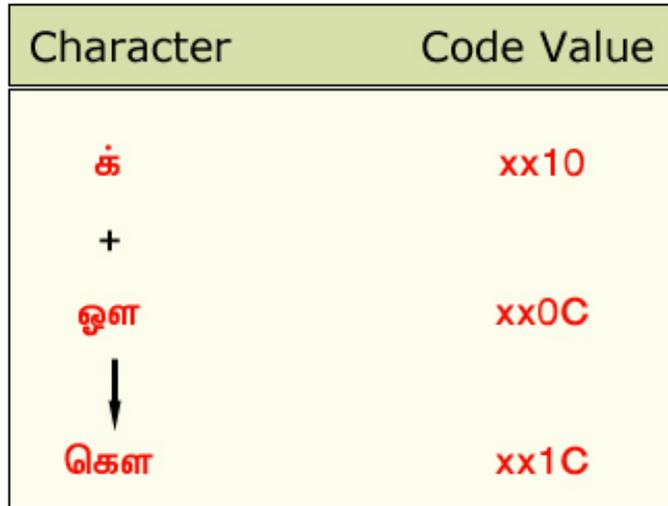
The old version has the code positions as given in **Table-1** below:

16-bit Tamil All Character Encoding (TACE_16)
16-பிட் தமிழ் அனைத்துரு குறியீட்டு முறை

	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	ண	த்	ந்	ப்	ம்	ய்	ர்	ல்	வ்	ழ்	ள்	ற	ன்	ஜ்	ஸ்	ஷ்	ஸ்த்	க்ஷ்	ஃ	ஃ	ஃ	ஃ
1	ஃ	க்	ங்	ச்	ஞ்	ட்	ண	த்	ந்	ப்	ம்	ய்	ர்	ல்	வ்	ழ்	ள்	ற	ன்	ஜ்	ஸ்	ஷ்	ஸ்த்	க்ஷ்	ஃ	ஃ	ஃ	ஃ
2	ஆ	கா	நா	சா	ஞா	டா	ணா	தா	நா	பா	மா	யா	ரா	லா	வா	ழா	ளா	றா	னா	ஜா	ஸா	ஷா	ஸா	ஹா	க்ஷா	ஃ	ஃ	ஃ
3	இ	கி	நி	சி	ஞி	டி	ணி	தி	நி	பி	மி	யி	ரி	லி	வி	ழி	ளி	றி	னி	ஜி	ஸி	ஷி	ஸி	ஹி	க்ஷி	ஃ	ஃ	ஃ
4	ஈ	கீ	நீ	சீ	ஞீ	டீ	ணீ	தீ	நீ	பீ	மீ	யீ	ரீ	லீ	வீ	ழீ	ளீ	றீ	னீ	ஜீ	ஸீ	ஷீ	ஸீ	ஹீ	க்ஷீ	ஃ	ஃ	ஃ
5	உ	கூ	நூ	சூ	ஞூ	டூ	ணூ	தூ	நூ	பூ	மூ	யூ	ரூ	லூ	வூ	ழூ	ளூ	றூ	னூ	ஜூ	ஸூ	ஷூ	ஸூ	ஹூ	க்ஷூ	ஃ	ஃ	ஃ
6	ஊ	கூ	நூ	சூ	ஞூ	டூ	ணூ	தூ	நூ	பூ	மூ	யூ	ரூ	லூ	வூ	ழூ	ளூ	றூ	னூ	ஜூ	ஸூ	ஷூ	ஸூ	ஹூ	க்ஷூ	ஃ	ஃ	ஃ
7	எ	கெ	நெ	செ	ஞெ	டெ	ணெ	தெ	நெ	பெ	மெ	யெ	ரெ	லெ	வெ	ழெ	ளெ	றெ	னெ	ஜெ	ஸெ	ஷெ	ஸெ	ஹெ	க்ஷெ	ஃ	ஃ	ஃ
8	ஏ	கே	நே	சே	ஞே	டே	ணே	தே	நே	பே	மே	யே	ரே	லே	வே	ழே	ளே	றே	னே	ஜே	ஸே	ஷே	ஸே	ஹே	க்ஷே	ஃ	ஃ	ஃ
9	ஐ	கை	நை	சை	ஞை	டை	ணை	தை	நை	பை	மை	யை	ரைய	லைய	வைய	ழைய	ளைய	றைய	னைய	ஜைய	ஸைய	ஷைய	ஸைய	ஹைய	க்ஷைய	ஃ	ஃ	ஃ
A	ஓ	கொ	நொ	சொ	ஞொ	டொ	ணொ	தொ	நொ	பொ	மொ	யொ	ரொ	லொ	வொ	ழொ	ளொ	றொ	னொ	ஜொ	ஸொ	ஷொ	ஸொ	ஹொ	க்ஷொ	ஃ	ஃ	ஃ
B	ஔ	கோ	நோ	சோ	ஞோ	டோ	ணோ	தோ	நோ	போ	மோ	யோ	ரோ	லோ	வோ	ழோ	ளோ	றோ	னோ	ஜோ	ஸோ	ஷோ	ஸோ	ஹோ	க்ஷோ	ஃ	ஃ	ஃ
C	ஓ	கொ	நொ	சொ	ஞொ	டொ	ணொ	தொ	நொ	பொ	மொ	யொ	ரொ	லொ	வொ	ழொ	ளொ	றொ	னொ	ஜொ	ஸொ	ஷொ	ஸொ	ஹொ	க்ஷொ	ஃ	ஃ	ஃ
D	ஃ																											
E																												
F																												

This system has the following advantages:

- ✓ The encoding is Universal since it encompasses all characters that are found in general Tamil text interchange.
- ✓ The encoding is very efficient to parse.



- ✓ By simple arithmetic operation the characters can be parsed

$$xx10 + xx0C = xx1C$$

$$\text{க்} + \text{ஓள} = \text{கௌ}$$

$$xx1C - xx10 = xx0C$$

$$\text{கௌ} - \text{க்} = \text{ஓள}$$

- ✓ 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.

But there are issues in implementing this scheme in the BMP area of the Unicode space. The issues are,

- There is no contiguous space in the BMP area to accommodate the entire character set of TACE16.
- If TACE16 is encoded in the BMP space, backward compatibility should be ensured for the present Unicode Tamil.
- It would be possible to place TACE16 in the SMP area of the Unicode space to reap the entire benefits of the proposed TACE16; But, this will increase the memory requirement of the Tamil contents when stored in the SMP space, since SMP is a 32 bit system, where as BMP is a 16-bit system.

4.2.2. Feasibility of Implementing the New TACE16

The new TACE16 has the code positions as given in **Table-2** below:

A suggested new code space for TACE-16 (18.7.2007)

	080	081	082	083	084	085	086	087	088	089	08A	08B	08C	08D	08E	08F	1CD	1CE	1CF	AA6	AA7	AAE	AAF	ABE	ABF
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																									
2	கா	நா	சா	ஞா	டா	ணா	தா	நா	பா	மா	யா	ரா	லா	வா	ழா		ளா	றா	னா	ஜா	ஸா	ஷா	ஸா	ஹா	க்ஷா
3	கி	நி	சி	ஞி	டி	ணி	தி	நி	பி	மி	யி	ரி	லி	வி	ழி		ளி	றி	னி	ஜி	ஸி	ஷி	ஸி	ஹி	க்ஷி
4	கீ	நீ	சீ	ஞீ	டீ	ணீ	தீ	நீ	பீ	மீ	யீ	ரீ	லீ	வீ	ழீ		ளீ	றீ	னீ	ஜீ	ஸீ	ஷீ	ஸீ	ஹீ	க்ஷீ
5	கு	நு	சு	ஞு	டு	ணு	து	நு	பு	மு	யு	ரு	லு	வு	ழு		ளு	று	னு	ஜு	ஸு	ஷு	ஸு	ஹு	க்ஷு
6	கூ	நூ	சூ	ஞூ	டூ	ணூ	தூ	நூ	பூ	மூ	யூ	ரூ	லூ	வூ	ழூ		ளூ	றூ	னூ	ஜூ	ஸூ	ஷூ	ஸூ	ஹூ	க்ஷூ
7	கெ	நெ	செ	ஞெ	டெ	ணெ	தெ	நெ	பெ	மெ	யெ	ரெ	லெ	வெ	ழெ		ளெ	றெ	னெ	ஜெ	ஸெ	ஷெ	ஸெ	ஹெ	க்ஷெ
8	கே	நே	சே	ஞே	டே	ணே	தே	நே	பே	மே	யே	ரே	லே	வே	ழே		ளே	றே	னே	ஜே	ஸே	ஷே	ஸே	ஹே	க்ஷே
9	கை	நை	சை	ஞை	டை	ணை	தை	நை	பை	மை	யை	ரைய	லைய	வைய	ழைய		ளைய	றைய	னைய	ஜைய	ஸைய	ஷைய	ஸைய	ஹைய	க்ஷைய
A	கொ	நொ	சொ	ஞொ	டொ	ணொ	தொ	நொ	பொ	மொ	யொ	ரொ	லொ	வொ	ழொ		ளொ	றொ	னொ	ஜொ	ஸொ	ஷொ	ஸொ	ஹொ	க்ஷொ
B	கோ	நோ	சோ	ஞோ	டோ	ணோ	தோ	நோ	போ	மோ	யோ	ரோ	லோ	வோ	ழோ		ளோ	றோ	னோ	ஜோ	ஸோ	ஷோ	ஸோ	ஹோ	க்ஷோ
C	கொ	நொ	சொ	ஞொ	டொ	ணொ	தொ	நொ	பொ	மொ	யொ	ரொ	லொ	வொ	ழொ		ளொ	றொ	னொ	ஜொ	ஸொ	ஷொ	ஸொ	ஹொ	க்ஷொ
D																									
E																									
F																									

The following are the issues in implementing this scheme in the BMP area of the Unicode space:

- The locations 0800 - 08FF are reserved for Samaritan, Mandaic and Arabic Extended-A, meant for right to left reading. These locations are used for the majority of characters in TACE16. Unicode Consortium need to be pressurized to release these locations from right to left limitation and assign for TACE. It should be possible since these locations are yet to be allocated to any other language.
- The TACE16 characters are placed in 6 different Blocks including Unicode Tamil Block. The vowels and agara-uyirmeys have higher code values (0B80 - 0BFF) than the consonants and other vowel-consonants (0800-08FF, 1CD-1CF, AA6F, AA7F, AAEF, AAF, ABEF, ABFF). This creates problems in morphological analysis, sorting and searching. This can be managed with a collation algorithm.
- The above problem will not exist if the Unicode consortium accepts to place the vowels in the locations 0801-080F and the agara-uyirmeys in the locations 0811-08F1, 1CD1, 1CE1, 1CF1, A61, A71, AAE1, AAF1, ABF1, ABF1, duplicating them both in the Unicode Tamil Block and in the TACE16 Block and call this as **modified New TACE16**. (Table-3)

A Modified new code space for TACE16

	080	081	082	083	084	085	086	087	088	089	08A	08B	08C	08D	08E	08F	1CD	1CE	1CF	AA6	AA7	AAE	AAF	ABE	ABF
0		க்	ங்	ச்	ஞ்	ட்	ண்	த்	ந்	ப்	ம்	ய்	ர்	ல்	வ்	ழ்	ள்	ற்	ன்	ஜ்	ஸ்	க்	ஸ்	ஹ்	க்ஷ்
1	அ	கா	ஙா	சா	ஞா	டா	ணா	தா	நா	பா	மா	யா	ரா	லா	வா	ழா	ளா	றா	னா	ஜா	ஸா	க்ஷா	ஸா	ஹா	க்ஷா
2	ஆ	கா	ஙா	சா	ஞா	டா	ணா	தா	நா	பா	மா	யா	ரா	லா	வா	ழா	ளா	றா	னா	ஜா	ஸா	க்ஷா	ஸா	ஹா	க்ஷா
3	இ	கி	ஙி	சி	ஞி	டி	ணி	தி	நி	பி	மி	யி	ரி	லி	வி	ழி	ளி	றி	னி	ஜி	ஸி	க்ஷி	ஸி	ஹி	க்ஷி
4	ஈ	கி	ஙி	சி	ஞி	டி	ணி	தி	நி	பி	மி	யி	ரி	லி	வி	ழி	ளி	றி	னி	ஜி	ஸி	க்ஷி	ஸி	ஹி	க்ஷி
5	உ	கு	ஙு	சு	ஞு	டு	ணு	து	நு	பு	மு	யு	ரு	லு	வு	ழு	ளு	று	னு	ஜு	ஸு	க்ஷு	ஸு	ஹு	க்ஷு
6	ஊ	கூ	ஙூ	சூ	ஞூ	டூ	ணூ	தூ	நூ	பூ	மூ	யூ	ரூ	லூ	வூ	ழூ	ளூ	றூ	னூ	ஜூ	ஸூ	க்ஷூ	ஸூ	ஹூ	க்ஷூ
7	எ	கெ	ஙெ	செ	ஞெ	டெ	ணெ	தெ	நெ	பெ	மெ	யெ	ரெ	லெ	வெ	ழெ	ளெ	றெ	னெ	ஜெ	ஸெ	க்ஷெ	ஸெ	ஹெ	க்ஷெ
8	ஏ	கே	ஙே	சே	ஞே	டே	ணே	தே	நே	பே	மே	யே	ரே	லே	வே	ழே	ளே	றே	னே	ஜே	ஸே	க்ஷே	ஸே	ஹே	க்ஷே
9	ஐ	கை	ஙை	சை	ஞை	டை	ணை	தை	நை	பை	மை	யை	ரையே	லையே	வையே	ழையே	ளையே	றையே	னையே	ஜையே	ஸையே	க்ஷையே	ஸையே	ஹையே	க்ஷையே
A	ஓ	கொ	ஙொ	சொ	ஞொ	டொ	ணொ	தொ	நொ	பொ	மொ	யொ	ரொ	லொ	வொ	ழொ	ளொ	றொ	னொ	ஜொ	ஸொ	க்ஷொ	ஸொ	ஹொ	க்ஷொ
B	ஔ	கோ	ஙோ	சோ	ஞோ	டோ	ணோ	தோ	நோ	போ	மோ	யோ	ரோ	லோ	வோ	ழோ	ளோ	றோ	னோ	ஜோ	ஸோ	க்ஷோ	ஸோ	ஹோ	க்ஷோ
C	ஓள	கொள	ஙொள	சொள	ஞொள	டொள	ணொள	தொள	நொள	பொள	மொள	யொள	ரொள	லொள	வொள	ழொள	ளொள	றொள	னொள	ஜொள	ஸொள	க்ஷொள	ஸொள	ஹொள	க்ஷொள
D	ஓ																								பூர்
E	ஓ	க	ங	ச	ஞ	ட	ண	த	ந	ப	ம	ய	ர	ல	வ	ழ	ள	ற	ந	ஜ	ஸ	க்ஷ	ஸ	ஹ	க்ஷ
F	0	க	ங	ச	ஞ	ட	ண	த	ந	ப	ம	ய	ர	ல	வ	ழ	ள	ற	ந	ஜ	ஸ	க்ஷ	ஸ	ஹ	க்ஷ

- If the above suggestion is accepted the benefit of parsing with simple arithmetic operations can employed for the **modified New TACE16** also as illustrated bellow:

$$\begin{aligned}
 & \text{க்} + \text{ஆ} = \text{கா} \\
 & 0810 + 0802 - 0800 = 0812 \\
 \\
 & \text{கா} - \text{க்} = \text{ஆ} \\
 & 0812 - 0810 + 0800 = 0802 \\
 \\
 & \text{ள்} + \text{ஓள} = \text{ளொள} \\
 & 1CD0 + 080C - 0800 = 1CDC \\
 \\
 & \text{ளொள} - \text{ள்} = \text{ஓள} \\
 & 1CDC - 1CD0 + 0800 = 080C \\
 \\
 & \text{ஜ்} + \text{இ} = \text{ஜி} \\
 & AA60 + 0803 - 0800 = AA63 \\
 \\
 & \text{ஜி} - \text{ஜ்} = \text{இ} \\
 & AA63 - AA60 + 0800 = 0803
 \end{aligned}$$

The only difference between the operations to be performed in the old TACE16 and the new TACE16 is to add or subtract a consonant 0800 which is a shift factor in the arithmetic operations.

5. STRATEGY FOR IMPLEMENTATION

- As could be seen from the above discussion, the best strategy for implementation would be to declare the old TACE16 as the State Standard and then move the Government of India to declare the same as the National standard. Then shall pursue the Unicode consortium to accept for incorporating the modified TACE16.

6. CONCLUSION

The Task force examined almost all the issues relating to the present Unicode Tamil and to the proposed alternative schemes through appropriate testing and evaluation procedures and through discussions with all concerned including the Unicode Consortium, some of the MNCs, INFITT, KaNithamizh Sangam and users in the Tamil Diaspora. Considering the outcome of the efforts made by the TASK FORCE, it is strongly recommended that the old TACE16 be announced as a 16-bit encoding Standard for Tamil by the Government of Tamil Nadu and be recommended to the Government of India for announcing the same as a National Standard.

7. ACKNOWLEDGEMENT

The Task Force is grateful to the Government of Tamil Nadu and acknowledges with gratitude for the opportunity given to the members for addressing an important issue of National interest. It places on record with gratitude the enormous support and help rendered by the testing agencies, the members of Unicode and all others concerned. The role played by the officials of TVU and the IT Department of the Government during the course of this exercise is commendable, without which it would not have been possible to complete the task entrusted to the Task Force and the acknowledges their services with utmost appreciation and gratitude.

ANNEXURE-1

Issues with the present Unicode Tamil

The present Unicode standard for Tamil is considered not adequate for efficient and effective usage of Tamil in computers, due to the following reasons:

- Unicode code Tamil has code positions only for 31 out of 247 Tamil Characters. These 31 characters include 12 vowels, 18 agara-uyirmey and one aytham. Five Grantha agara-uyirmey are also provided code space in Unicode Tamil. The other Tamil Characters have to be rendered using a separate software. 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.
- The Uyir-meys that are left out in the present Unicode Tamil are simple characters, just like A, B, C, D are characters to English. Uyir-meys are not glyphs, nor ligatures, nor conjunct characters as assumed in Unicode. ka, kA, ki, kI, etc., are characters to Tamil.
- In any plain Tamil text, Vowel Consonants (uyir-meys) form 64 to 70%; Vowels (uyir) form 5 to 6% and Consonants (meys) form 25 to 30%. Breaking high frequency letters like vowel-consonants into glyphs is highly inefficient.
- This type of encoding which requires a rendering engine to realize a character while computing is not suitable for applications like system software developments in Tamil, searching and sorting and Natural language processing in Tamil, It consumes extra time and space, making the computing process highly inefficient. For such applications Level-1 implementation where all the characters of a language have code positions in the encoding, like English is required.
- This encoding is based on ISCII - 1988 and therefore, the characters are not in the natural order of sequence. It requires a complex collation algorithm for arranging them in the natural order of sequence.
- It uses multiple code points to render single characters. Multiple code points lead to security vulnerabilities, ambiguous combinations and requires the use of normalization.
- Simple counting letters, sorting, searching are inefficient
- It requires ZWJ/ZWNJ type hidden chars.
- It needs exception table to prevent illegal combinations of code points.

- Unicode Indic block is built on enormous, complex, error-prone edifice, based on an encoding that is NOT built to last.
- Very first code point says “Tamil Sign Anusvara - Not used in Tamil”.
- Assumed collation was same as Devanagari - incorrectly uses ambiguous encoding to render same character.
- It encodes 23 Vowel-Consonants (23 consonants + அ) and calls them as consonants, against Tamil grammar.
- Unnatural for Speech to Text/Text to Speech.
- Inefficient to store, transmit and retrieval.
- Complex processing hinders development.
- Need normalization for string comparison.
- A sequence of characters may correspond to a single glyph, that is, ச + ெ + ா = சொ. Characters are not graphemes. According to Unicode சொ is a grapheme; but ச, ெ, ா are characters.
- Requires Dynamic Composition - a text element encoded as a sequence of a base character followed by one or more combining marks.
- There are two methods of rendering the following class of Vowel Consonants.

$$\begin{aligned}
 \text{ச} + \text{ொ} &= \text{சொ} = \text{ச} + \text{ெ} + \text{ா} \\
 \text{ச} + \text{ோ} &= \text{சோ} = \text{ச} + \text{ே} + \text{ா} \\
 \text{ச} + \text{ெ்ள} &= \text{செள} = \text{ச} + \text{ெ} + \text{்ள}
 \end{aligned}$$

This leads to ambiguity in rendering characters.

- The Present Unicode is not efficient for parsing. For example, let us count the letters in the name திருவள்ளுவர். Even a Tamil child in a primary school can say that this name has Seven letters. According to Unicode this name has twelve characters:

த ி ர ு வ ள ன் ள ு வ ர ்

- To properly count the letters in this name, an expert developer had to write a complex program and present it as a technical paper in a Tamil computing

conference¹. To compare, counting letters in an English word is an exercise left to a beginning programmer. Such problems are triggered because a simple script such as Tamil is treated as a complex script by Unicode.

- The Unicode standard policy is to encode only characters, not glyphs. However, because Unicode Tamil standard includes the following vowel signs as combining characters

ா ி ு றெ ோ ொ றெள

these signs that have no meaning to a Tamil reader would be displayed as is by character shaping engines that detect a blank space between them and a base character. Thus Unicode introduces the dotted circle as a Tamil character!

- Unicode Tamil is not fully supported in many platforms primarily because Tamil is treated as a complex script that requires complex processing.

¹ “Counting Letters in an Unicode String” (sic), T. N. C. Venkatarangan, Presented at Tamil Internet Conference 2004, Singapore

**Results of the investigation of the 16-bit All Character Encoding for
Tamil carried out in 2001**

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



வட அமெரிக்கத் தமிழ்ச்சங்கப் பேரவை
Federation of Tamil Sangams of North America
(A Registered, Non-Profit, Tax - Exempt Organization EIN# 36-3759132)

2756 Sunbonnet Court, San Jose, CA 95125
www.fetna.org

தேமதுர தமிழோசை உலகமெல்லாம்
பரவும் வகை செய்தல் வேண்டும்

President

Thiru Thillai K. Kumaran 408 267 8006

Vice President

Dr. R Probhakaran 410 420 0111

Secretary

Thiru Mani Kumaran 703 378 6094

Joint Secretary

Thiru Ravi Shanmugam 919 465 1606

Treasurer

Thiru. Sri Subramaniam 863 421 7949

Directors

Thiru. S. Sittambalam 416 444 4931

Thiru. Pazhani Sundaram 203 271 2064

Thiru. O.P. Ravanan 732 985 1347

Thiru. R Porchezhan 314 878 5832

Tamil Sangams

Austin Tamil Sangam
San Francisco Bay Area Tamil Manram
Bharathi Kalai Manram, Houston
Bharathi Society of America, New York
Boston Muthamizh Sangam
California Muthamizh Sangam
Chicago Tamil Sangam
Colorado Tamil association
Columbus Tamil Sangam
Connecticut Tamil Sangam
Greater Atlanta Tamil Sangam
Greater Charlotte Tamil Association
Harrisburg Area Tamil Sangam
Iankai Tamil Sangam of Florida
Iankai Tamil Sangam of Texas
Iankai Tamil Sangam of USA
Indianapolis Tamil Sangam
Metroplex Tamil Sangam
Midwest Thamizh Sangam
Milwaukee Tamil Sangam
Minnesota Tamil Sangam
Muthamil Sangam, Orlando
National Capital Tamil Service Society
New England Tamil Sangam
New Jersey Tamil Arts & Cultural Society
New Jersey Tamil Sangam
New Mexico Tamil Sangam
New York Tamil Sangam
Panai Nilam Tamil sangam
Oklahoma Tamil Sangam
San Antonia Tamil Sangam
San Diego Tamil Sangam
Seattle Tamil Sangam
South Florida Tamil Sangam
Tamil Association of Greater Delaware
Tamil Eelam Society of Canada
Tamil Manram, Jacksonville
Tamil Sangam of Carolinas
Tamil Sangam of Greater Washington
Tamil Sangam of Los Angeles
Tampa Tamil Association of USA
Tennessee Tamil Sangam
Thamizh Sangam of Missouri
Wisconsin Tamil Sangam

July 3, 2006 – New York

Resolution

ON THE CRISIS OF TAMIL COMPUTING ENCODING

WHEREAS, The crisis of Tamil computing encoding impacts and will continue to impact millions of Tamil speaking people in many different ways; and

WHEREAS, Tamils have responsibilities for the continuity and effective use of the Tamil language using modern technology; and

WHEREAS, As part of the Indian nation, the Tamil Nadu government is under constitutional mandate to respect the national standards and the international standards that the Government of India recognizes, but at the same time, Tamils have a right to expect the both the State and Union Governments to address the unique needs of the Tamil language and its speakers throughout the world; and

WHEREAS, as users of modern technologies for information storage and communication in the Tamil language we are concerned about the correct application of the technology for Tamil language; and

WHEREAS, the Tamil Unicode segment as it stands today is designed primarily for the Devanagari based Indic languages; and

WHEREAS, the Tamil Nadu Government, through the agency of Tamil Virtual University, has created a better encoding for Tamil using the all character encoding known as TUNE (Tamil Unicode New Encoding);

WHEREAS, this TUNE encoding needs the endorsement and support of the Union Government of India; and

WHEREAS, The union government's inability to resolve the Tamil encoding crisis amicably, has caused severe consternation among the users of Tamil language on the internet and computers and has led to the crisis we now face; and

WHEREAS, The union and state government have not only failed to agree on a computer encoding standard for Tamil but failed to enforce the various competing standards in government documents, archives as well as Tamil based internet web pages in India; and

WHEREAS, The governments have failed to adequately represent the concerns of the Tamils to the Unicode Consortium and other international bodies and multinational corporations; and

WHEREAS, These encoding standards will have a great impact on the archives of billions of government records that are being digitized; now, therefore, be it

RESOLVED, That the delegates to the Federation of Tamil Associations of North America Convention meeting in New York, New York, July 3, 2006, urge the Union Government of India and the State Government of Tamil Nadu, to recognize the TUNE encoding as the standard 16-bit encoding for Tamil Language; and be it further

RESOLVED, That we urge the Union Government of India and the State Government of Tamil Nadu to enforce the TUNE encoding as the Indian national standard for Tamil 16-bit encoding in a way that will restore trust among the Tamil speakers; and be it further

RESOLVED, That we urge the Union government of India to enforce the TUNE encoding in all the Tamil software that is sold to the Central and State Governments of India; and be it further

RESOLVED, That we urge Tamil language users on computers and internet to follow the TUNE encoding as a standard in their use of Tamil in computers.

Sincerely,

Thillai K Kumaran
President

எங்கள் வாழ்வும் எங்கள் வளமும் மங்காத
தமிழென்று சங்கே முழங்கு



ABSTRACT

Information Technology – Tamil Virtual University - Conference on Tamil 16-bit All Character Encoding System held on 2.9.2006 – Adoption of the recommendations made in the Conference-Constitution of Task Force and entrusting work to formulate action plan for implementation of the recommendations – Orders – Issued.

INFORMATION TECHNOLOGY DEPARTMENT

G.O.(Ms) No.13

Dated:10.11.2006

Read:

G.O.Ms.No.17, IT Department, dated 13.06.1999

ORDER

In the G.O. read above, orders were issued announcing a unique 8-bit Tamil encoding system. Soon after these standards were announced, a number of application software systems in Tamil, like Tamil word processors, Tamil OCR, Tamil search engines, etc., were developed. However, these 8-bit encoding schemes could not provide code positions for all Tamil characters in the encoding chart due to the limitation in space (Total available space is only 128 locations).

2. Realizing the limitations of the 8-bit encoding and the present 16-bit Unicode Tamil, the Government, in 1999 itself, announced, at the time of declaring 8-bit encoding standard for Tamil, that an efficient 16-bit character encoding will be developed for Tamil and will be submitted to the Unicode consortium for incorporation in the Unicode standard. Accordingly, the Government initiated action in the direction through the Tamil Virtual University (TVU). The TVU formed a committee with experts, pooled from KaNithamizh Sangam for this purpose. The Committee developed an all Character 16-bit encoding scheme for Tamil. The proposed scheme was presented at the TamilNet conferences held during the years 2000,2001,2002 and 2003 in Sri Lanka, Singapore, Malaysia, California (USA) and Chennai.

3. Meantime the Government have initiated action through the TVU for the submission of the new scheme to the Unicode consortium for consideration. In order to facilitate the submission of a revised character encoding standard for Tamil for incorporation into Unicode, the Government became an Associate Member in the Unicode Consortium in 1999. Afterwards, series of actions were taken by the TVU for testing the proposed scheme and processing the same for incorporation in the Unicode Standard.

4. The Chairman, TVU has stated that the testing of the all character encoding scheme for Tamil, containing 24 X16 code points was completed and the results were uploaded in the Government website requesting for comments from the Tamil Diaspora. The Government submitted the above scheme to the MIT, GOI and to the Unicode Consortium for consideration. The Unicode Consortium suggested to put the proposed all character encoding scheme in the Private Use Area (PUA) of Unicode, so as to put the scheme in use by the Tamil Diaspora. Accordingly a new keyboard driver and a new Unicode font for the new encoding were developed in order to test the new encoding in PUA under various operating platforms and in all possible applications. The all character new encoding was thus placed in the Unicode block E200 to E38F on 24th June 2005.

5. A conference on Tamil 16-bit All Character Encoding was organized on 2nd September 2006 by the TVU to review the comments received and to plan further course of action to declare one single true 16-bit all character encoding scheme for Tamil as a standard in the place of existing 8-bit encoding standards and for moving the same into Unicode. The conference was presided over by Prof. V.C.Kulandaiswamy, Chairman, TVU and inaugurated by Thiru Dhayanithi Maran, the Hon'ble Minister for Communications and Information Technology, Government of India. The Conference was attended to by the delegates from Singapore and Sri Lanka. Software professionals from major firms such as IBM, MICROSOFT and Tamil Software developers in Tamil Nadu and in the other parts of the country and abroad participated. In his inaugural address, the Hon'ble Minister said that the new scheme shall be reviewed and revised based on the comments received and tested on different platforms and in different applications like E-Governance, web publication, etc. The Hon'ble Minister further said that creation of a corpus fund will be considered for testing and development and for encouraging migration and conversion to the new encoding. Hon'ble Minister desired that we should work with a united mind and on a time bound programme, so that the 16-bit encoding for Tamil is available for implementation in the e-Governance project in January 2007.

6. Following the inaugural session, a Technical session for open discussion for the delegates to express their views on Tamil 16 Bit All character Encoding system was held on 2.9.2006. The Session was chaired by Dr.V.C.Kulandaiswamy and a report on the details of the activities towards the development of the new encoding scheme was presented by Dr.M.Ponnaivaikko, former Director, TVU. The delegates participated actively in the discussion and gave useful suggestions. The Technical Committee constituted by the TVU with Dr.M. Anandakrishnan as Chairman, Dr. M. Ponnaivaikko as convener,

Dr.N.Balakrishnan, Associate Director, IISc., Bangalore, Dr.M.N. Cooper of Modular Infotech., Tmt. Swaran Latha, Scientist E, MCIT and a few other Experts as special invitees deliberated on the issue in the afternoon Session. 80% of the participants voted for the expeditious implementation of the 16-bit Tamil All Character Encoding as unique encoding standard for Tamil.

7. In the Conference held on 2.9.2006 the following recommendations were made:-

(i) Recommended to consider the formation of Task Force to co-ordinate the activities related to the development of an acceptable 16-bit All Character Encoding for Tamil Language through appropriate testing and validation with the following mandates;

- a) To finalize the 16-bit Tamil All Character system of encoding by end December 2006;
- b) To test extensively the 16-bit Tamil All Character scheme of encoding for critical major applications such as e-Governance, web publishing, etc;
- c) To disseminate the proposed 16-bit Tamil All Character encoding extensively worldwide, and
- d) To convene an International Conference for obtaining the final consensus on the 16-bit Tamil All Character encoding from the Tamil Diaspora so that implementation of 16-bit Tamil All Character scheme of encoding may be considered, if the consensus is in favour of the 16-bit Tamil All Character encoding.

(ii) To take action to publicize the proposed 16-bit Tamil All Character scheme of encoding in the countries where Tamil is an official language so as to get their comments on the proposed scheme;

(iii) To create a corpus fund for providing financial and policy support for migrating contents and developments already done in the current environment;

(iv) The corpus fund created may include funds for developing tools and drivers to support the 16-bit Tamil All Character encoding in different platforms such as windows, Macintosh, Linux and Unix for free distribution; and

(v) The Government of Tamil Nadu to become a full voting member of Unicode Consortium, so that the State can directly submit proposals to Unicode consortium for adopting Tamil-16 bit All Character encoding into Unicode.

8. The Govt. after careful examination have decided to accept and to adopt the recommendations made in the Conference referred to in paragraph 7 above.

9. The Government, accordingly, constitute a Task Force with the following members:-

Task Force

1. Chairman : Dr. M.Anandakrishnan
Chairperson, Madras Institute of Development Studies,
No.79, Second Main Road,
Gandhi Nagar, Adyar, Chennai-20
2. Vice Chairman : Dr.M.Ponnaivaikko,
Director,
Research and Virtual Education,
SRM Institute of Science & Technology,
SRM Nagar, Kattankulathur,
Kanchipuram District, Pin - 603 203.
3. Members : (1) Dr. N.Balakrishnan,
Associate Director,
Indian Institute of Science,
Bangalore- 560 012.
- (2) Dr. M.Rajendiran,
Director, Directorate of Tamil
Development
Tamil Development Campus,
Near to Children's Hospital,
Halls Road, Egmore, Chennai- 600 008.

- 3) Thiru.A.Mohan,
Deputy Director General,
National Informatic Centre,
Rajaji Bhavan, Besant Nagar,
Chennai-600090
- 4). Tmt. Swaran Latha,
Scientist -E,
Ministry of Communication &
Information Technology, New Delhi
- 5). Dr.S.Ramakrishnan,
Executive Director,
Centre for Development Advanced
Computing(CDAC), Pune
- (6) Dr. M.N. Cooper,
Managing Director,
Modular Infotech Limited,
26, Electronic Co-operative Estate,
Pune-Satara Road, Pune - 411 099.
- (7) Mr. P.Chellappan,
Former President-KaNithamizh Sangam,
Palaniappa Brothers,
No.14, Peters Road,
Chennai- 600 014.
- (8) Dr. V.Krishnamoorthy,
Professor, Crescent Engineering College,
A2, Navaneeth Flats, New No.13,
Venkatarathinam Nagar Extn. - II Street,
Adyar, Chennai- 600 020.
- (9) Mr. M.Anto Peter,
President, KaNitamizh Sangam,
Softview Computers,
118, Nelson Manickam Road,
Metha Nagar, Chennai-600 042.
- (10) Mr. N.Anbarasan,
C.E.O. Applesoft,
39, 1st Cross 1st Main
Shivanagar, W.C. Road,
Bangalore-560 010.

4. Convenor : Dr. P.R. Nakkeeran
Director,
Tamil Virtual University,
4th Floor, ELNET Software City,
CPT Road, Taramani,
Chennai - 600 113.

10. The Government direct that the Task Force shall formulate action plan for the implementation of the following recommendations made in the conference held on 2.9.2006:-

- (i) Action to publicize the proposed 16-bit Tamil All Character scheme of encoding in the countries, where Tamil is an official language so as to get their comments on the proposed scheme;
- (ii) To create a corpus fund for providing financial and policy support for migrating contents and developments already done in the current environment;
- (iii) The corpus fund to be created shall include funds for developing tools and drivers to support the 16-bit Tamil All Character encoding in different platforms, such as windows, Macintosh, Linux and Unix for free distribution; and
- (iv) The Government of Tamil Nadu shall become a full voting member of Unicode Consortium, so that the State can directly submit proposals to Unicode consortium for adopting Tamil-16 bit All Character encoding into Unicode.

11. The Chairman, TVU has recommended that the following persons in overseas may also be considered for inclusion in the Task force so as to have an International representation:-

(1) Mr. Mani Manivannan, California, USA
mmanivannan@gamil.com

(2) Dr. K. Kalyanasundaram,
RUE DAS APLES 59,
1030, BUSSIGNY, Switzerland,
kalyan.geo@yahoo.com

(3) Dr. Kalaimani, Singapore

The Government accept the recommendation of the Chairman TVU. They accordingly request the Chairman of the Task Force to seek the advice of the above overseas experts over email/correspondence if he so desires.

12. The Chairman, Task Force is requested to take necessary follow-up action in this regard and send a report to the Government at an early date.

(BY ORDER OF THE GOVERNOR)

**C.CHANDRAMOULI
SECRETARY TO GOVERNMENT**

To

The Chairman, Tamil Virtual University, Chennai-113.

The Director, Tamil Virtual University, Chennai-113.

The Managing Director, ELCOT, Chennai-35.

The OSD, e-Governance Directorate, Chennai-35.

All Members of Task Force

Copy to:

All Secretary's to Government, Chennai-9

Chief Minister's Office, Chennai-9

Finance (BPE) Department, Chennai-9

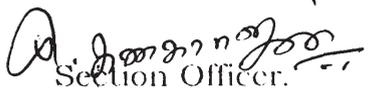
All Sections in IT Department, Chennai-9.

Private Secretary to Chief Secretary, Chennai-9.

Private Secretary to Secretary, IT Department, Chennai-9.

SF/SC

/Forwarded / By Order /


Section Officer.

ANNEXURE-5

Formation of Subcommittee to Examine the encoding of Tamil and other scripts of India

(Mail from Rick McGowan of Unicode, Inc.)

Welcome to the South Asian subcommittee mail list.

The subcommittee has been formed by UTC to examine the encoding of Tamil and other scripts of India. To send mail to the list, you may address it to "southasia@unicode.org".

Eric Muller is the subcommittee chair.

Unicode members are invited to subscribe if they wish, by sending mail to "ecartis@unicode.org" with "subscribe southasia" in the subject line. Topic areas include scripts and languages of the Indian subcontinent, including scripts of India, Pakistan, Bangladesh, Nepal, and Sri Lanka. (Excludes Arabic-based scripts, which are handled by the Bidi Subcommittee.)

The following people have been subscribed to this list initially:

Pankaj Agrawala (Gov't of India)
Debbie Anderson (U C Berkeley)
Manoj Annadurai (Gov't of India)
Lee Collins (Apple)
Peter Constable (Microsoft)
Somdutt Dadheech (Gov't of India)
Mark Davis (Google)
Deborah Goldsmith (Apple)
Cibu Johny (Google)
Michael Kaplan (Microsoft)
Mani Manivannan (Gov't of Tamil Nadu)
Rick McGowan (Unicode)
Ram Mohan (Afilias)
Lisa Moore (IBM)
Muthu Nedumaran (Sponsored by Apple)
M. Ponnaivaikko (Gov't of Tamil Nadu)
Michel Suignard (Microsoft)
Tex Texin (Yahoo)
V S Umamaheswaran (IBM)
Ken Whistler (Sybase)

I will announce locations for documents and Wiki information at a later date, as the infrastructure becomes available.

Regards,
Rick McGowan
Unicode, Inc.

Report of the South Asia Subcommittee Meeting in Chennai, 2008/01/23-24

The following is a report on the South Asia Subcommittee Meeting in Chennai.

Initial discussions

The following are informal notes that summarize the address from the IT Secretary of the Tamil Nadu Government. The exact words will be attached later.

The government is concerned about some things:

1. Errors in encoding such as non-Tamil characters being in the encoding
2. Efficiencies - Govt of Tamil Nadu is undertaking a massive e-governance efforts. Huge digital libraries are coming up and govt doesn't want to migrate these massive databases in the future. Govt relies on the experts in the task force and 13 meetings have been held to review and analyze TACE encoding.
3. There may be legal issues as well and the government has to be very careful.

The government's position is that one block for TACE-16 in Unicode would be desirable. The feasibility and practicality need to be investigated. Urged UTC /UC to look at and suggest ways to resolve the issues. The problems are genuine. Tamil is an international language. Used in official transaction in Sri Lanka. International ramifications. We have proposed our solution. We would like UTC's recommendations.

As for Tamil Nadu Government, it intends to accept the recommendations of the task force and declare a standard and expects the Government of India's support as well.

- We do appreciate that a lot of work has been done.
- Appreciate everyone's being as frank and forthcoming as possible.
- What are the objections, stance of the UC.
- Also need to consider the Tamil Community's stance
- The UC's views are vital, and can suggest better solutions if available
- We would like international community's opinion to make an informed decision
- Announced creation of a fund to ease the migration from the old to new.
- Some teething problem from old to new
- Migration path can be considered.

The government of TN wants to hold a conference in coordination with INFITT.

Opening remarks from Mark Davis; main points in summary:

- The UC also has the goal of making Tamil work correctly; we look forward to working with the TNG and the GOI to do this.
- An important issue is stability - Unicode is a bit like the banking system - the confidence of members and implementers in the stability is key to its success.
- Not breaking existing implementations is vital

Discussion

- General agreement about the need to improve the situation for Tamil Users
- Key problem is the lack of implementations of Tamil, and the correctness of those implementations. There were many examples of the need to improve.
- For the meeting, we'll be following the South Asia subcommittee charter as per the August UTC meeting, copied below.

Step 1: Identification of the issues

Discussion of TACE-16 - proposal for 347 new characters replacing Tamil block. Following issues were raised as part of that presentation.

- Significant work since the May meeting on identifying the issues, developing concrete tests
- The task force on TACE-16, which are technical advisors to the Government of Tamil Nadu, has recommended to the government the All Character Tamil encoding (TACE-16) be made a government standard to meet the following requirements:
 - Handle the emerging needs of e-governance,
 - Produce unambiguous and legally indisputable digital records of government documents
 - Enable creation of documents that will stand the test of time,
 - Be Independent from any external shaping engine
 - Be efficient in desktop publishing, linguistic and natural language processing
 - Assure safe, unambiguous browsing resistant to domain name spoofing
- Issues Raised
 - What users think of as characters (syllabic)
 - Ambiguous encodings in Unicode (length marks)
 - Unicode characters not in collation order
 - Simplification of natural-language processing
 - Dependence on correct rendering engines
 - Fonts not having correct OpenType tables
 - Variable levels of support for OpenType fonts
 - Efficiency in storage and processing
 - *See attached report for more details.*
- Conclusion by TACE task force that Unicode does not match the user's perceptions of characters, and is less optimal than TACE-16 for the measured operations.

In discussion, an issue was raised:

- the results need to be reproducible: eg, data made available and at least pseudo-code for the operations

Discussion of Unicode principles relevant to above (Mark Davis, Michael Kaplan)

1. Unicode character is coded entity ≠ what user thinks of as "letter" or "character". Many examples from variety of scripts. This is true for languages other than Tamil as well. (e.g. Swedish A with a ring character).
2. Canonical equivalence establishes identity; normalization (NFC) used for unambiguous representation (specifically, the "broken" vowel pieces are combined). Used in important cases like IDNA.
3. Code order ≠ collation order for any language: eg, Z < a
4. Display ≠ character codes. Many scripts require more than linear layout. Some of the errors or inefficiencies may be triggered by problems in correctness of implementations. For example, collation, rendering, etc. OpenType fonts with ligature tables for all of the 345 or so Tamil characters identified could be precomposed and mapped to existing Unicode quite efficiently.
5. Storage is an issue, but not predominant (discussion of UTFs, history of UTF-16). (See also #9)
6. Stability is a key issue. Unicode is like the banking system. People have to be able to trust that it won't change out from under them. Major clients of Unicode are very dependant on this -- some would much rather have stability than improvements.
7. Similarity of models helps with implementation. Perceived difficulty of implementation only increases if the language deviates from a family model and stands by itself. There is strength in being part of a family model where only slight modifications are needed to support a new language. (For example, other Indic scripts.)
8. There may be minority letters in script, or "mistaken" characters like [U+0B82](#) (◌◌) TAMIL SIGN ANUSVARA which is not used in any language using the Tamil script. Characters can be annotated (as Anusvara is), or deprecated (stronger), but never removed. (There was discussion of options for this character, as to whether to annotate or deprecate.) Even the name of the character cannot be changed. (There are separate data files in the Unicode Character Database with name annotations, more information, and with correction.) Note: localized names for Tamil characters can be supplied (eg Pulli vs "VIRAMA", or visarga), so that vendors can display the correct name in programs like CharMap.
9. Results of efficiency vary dramatically according to the code used. Efficiency in storage/transmission are implementation dependent and algorithms can be carefully optimized. Efficiency in processing is a desirable goal but if stability of implementation forces a hit on efficiency that is acceptable. (See also #5)

Notes:

SMP vs BMP

1. BMP is from 0000 to FFFF. Most common characters, widely supported
2. SMP is from 10000 to 10FFFF. Infrequent characters, historic scripts. Support in major OS's began a few years ago, but many applications don't fully support. (Examples: Vista supports plane 1 and 2 (only fonts for plane 2)).

3. BMP code points are typically transmitted as 3 UTF-8 octets while SMP requires 4. In UTF-16, these are 2 bytes for BMP characters, and 4 for SMP characters. (The difference is not double as might be expected.)
4. Space in the SMP is not constrained, whereas space in the BMP is very confined at this point. In particular, certain areas are reserved for Right-Left characters, which cannot be changed without serious consequences.

Discussion

- Members of the TACE taskforce disputed the points about the efficiency/performance issues, and benefits of following the Indic model.
- At the time that Tamil was first encoded, it could have followed a syllabic model for encoding like Ethiopic has now.
- Implementations quite often may transform Unicode into different internal formats for processing, such as in doing natural-language processing.
- If TACE were in the SMP, some problems are avoided -- the main blocker is dual encoding and stability.
- Normalization cannot map old characters to new characters, for stability constraints. If a new precomposed character were added, then it would normalized back to its components.
- Unicode operating systems (Windows, Mac, etc) convert to Unicode for rendering, etc.

Step 2: Evaluation of possible approaches

We started from the bottom up:

Approach D: TACE-16 as a separate IANA-registered character set
Unicode programs would convert on input to Unicode, process, and emit TACE-16 on output. (Similar to GB 18030.) Non-Unicode programs could process natively.

Pros

1. *No dependency on Unicode - Tamil Nadu government can do independently*
2. TACE16 is very easy to implement; not stateful, easy conversion to and from Unicode
3. well-established path for charsets -implementations are used to using them
4. governments have strong sway
5. the Tamil Nadu government can do exactly what it wants
6. useful in any closed environment: examples: cell phone, natural-language processing, etc.
7. well-defined path for programs to support -- programs are used to doing conversion
8. if multilingual capabilities are required *inside the same codepage*, then additional repertoire would need to be, eg, for English, French, Telugu, Malayalam, Sinhala, etc.
 1. Example: GB 18030 (China) includes all Unicode characters, with an algorithmic mapping to Unicode for most characters.

2. The simpler the mapping to Unicode, the more likely implementations would pick it up.
9. See iana.org for the list of IANA charsets.
10. Other TACE advantages: eg Processing using syllables (eg NLP) would use single code points.
11. On Unicode system, where conversion is done, algorithms depending on Unicode properties would work: line-breaking, sorting, identifiers, etc.

Cons

1. whether it is added to products depends on company's adding the conversion tables.
 2. for cell-phone environments, 8-bit encoding may be preferred
 3. uptake by companies will depend on critical mass, so a bit of a chicken and egg problem
 4. performance issues need investigation
 5. Typically Unicode programs / OSs will convert to Unicode for rendering, etc. (Linux may not -- needs investigation.) However, typically performance is not substantially impacted for rendering.
 6. Would need to evangelize key players
 1. ICU, Windows, Java, PHP, Python, Perl, Linux,...
 2. Many will pick up without further evangelization
 3. Most are combination of data+algorithms
-

Approach C: TACE-16 repertoire in the PUA

Pros

1. *No dependency on Unicode - Tamil Nadu government can do independently*
2. Encapsulated in Unicode, so no conversion necessary
3. SMP PUA is unencumbered - TACE-16 group could establish precedent (homesteading)
4. Compression of SMP works well
5. Rendering would be straightforward.
6. Other TACE advantages: eg Processing using syllables (eg NLP) would use single code points.

Cons

1. BMP PUA is in wide use for ideographs already, so it probably wouldn't be practical. (needs investigation, there might be enough room)
2. Overlap problem - some others could use code points for different purpose
3. Many implementations, & all old implementations, will treat as unknown characters (impacting anything dependant on properties: line-breaking, sorting, identifiers, etc). No standard Unicode properties, so algorithms driven by them won't work

4. Conversions are needed for interfacing with standards that require standard Unicode. For example, IDNs will be in standard Unicode, requiring a conversion.

Discussion:

- legal implications of PUA:
 - If the Tamil Nadu government established a standard, then being a standard for legal purposes is not an issue.
 - For legal purposes, people need to use final-form document with embedded fonts, for any language.
 - Font issues are not specific to PUA - can have font-spoofing in either way.
-

Approach CD: Approach C, plus register it with IANA as a charset.

1. Mixture of advantages and disadvantages of above.
 2. Examples:
 1. *No dependency on Unicode - Tamil Nadu government can do independently*
 2. In some cases, TACE would convert to and from Unicode; in others it could be interpreted natively.
 3. Character properties would be available; all multilingual capabilities would be present;
 4. IANA pros and cons from D.
-

Approach B: TACE-16 repertoire added to Unicode

TACE-16 task force investigated different approaches (listed above, and with full report attached). Major choices are BMP vs SMP. The TACE task force would like to see TACE in the BMP; failing that, the SMP would be an acceptable backup.

Pros

1. See attached document

Cons

1. Unless current Unicode model can be shown to not be able to represent Tamil, the duplicate encoding and stability principles would prevent addition.
2. Accommodating TACE in the BMP would require moving the reserved RTL (U+0800 .. U+08FF) code point range. (Space is not an issue for the SMP.)
 - The suggestion from the TACE group is to move the reserved RTL area to

1. Arabic extensions to U+18B0 .. U+18FF
 2. Mandaic to U+A8E0 .. U+A8FF
 3. Samaritan to U+AB50 .. U+AB7F
 4. Sorang Sng to U+A4D0 .. U+A4FF
-

Approach B1: Add only "pure consonants" to Unicode

This would be adding what is currently represented as <consonant + pulli> as precomposed characters to the current Tamil block in Unicode.

Pros

1. Pure consonants represent 30% of the letter frequency in Tamil text
2. Possible performance benefits in collation, text size (for unnormalized text)

Cons

1. Would be introducing new precomposed characters
 2. Normalization would replace the new characters with the current ones.
-

Key Areas where governments, industry, and Unicode can help

There is a natural frustration with programs not being able to handle Tamil, or having errors. Discussed common techniques companies use in prioritizing their work on different languages, and how to leverage improvements.

No matter what approach is taken, common need for the following (draft list)

1. Identify problems in key application programs and set up communication with vendors
2. Core set of open source (individual *and* commercial use) high-quality fonts
3. Freely available keyboard specifications and IMEs
4. Central place for developers to go for help with Tamil (on Unicode site or Government site, perhaps wiki?)
5. Up-to-date locale data (eg CLDR)
6. Need to investigate having standard ligature table for OpenType to map Unicode sequences to TACE syllables.

Side issue: the Tamil numbers are almost archaic, and offer opportunities for spoofing, so are discouraged for identifiers such as IDN.

Discussion of Unicode Locales Project (CLDR)

- (not able to do for lack of time)

We wish to thank our hosts, the Tamil Virtual University and Government of Tamil Nadu

South Asia Charter for Tamil Discussion (L2/07-272, item 10)

Goal: ensure that Unicode meets the needs for representation and processing of Tamil.

This may or may not require the encoding of new characters. Any recommendation should exhaustively examine the implications, including on existing data, on existing software (processing, display, etc), on education about the standard, on consistency of model for the Indic and other South Asian scripts.

The scope of the subcommittee is to review the issues and to make recommendations to the UTC.

Step 1: Identification of the issues Identify the issues (problems or perceived problems) with the current representation. Determine whether they are issues with the standard itself (encoding, properties, or algorithms) or with implementations. Determine the nature of the issues: technical, perceptual or educational.

Candidate issues:

- 1 disconnect of the code chart with the user expectations
- 2 efficiency in storage/transmission
- 3 efficiency in processing
- 4 correctness of implementations
- 5 difficulty of implementation

Step 2: Evaluation of possible approaches

This enumeration of possible approaches does not preclude the examination of other approaches (which may extend on or combine the approaches below). The questions listed for each approach are illustrative of the kinds of questions that need to be answered for a proper evaluation of the approach; they are not exhaustive.

Approach A: current model

How would those issues be addressed with the current representation? Are there any enhancements (new characters, changes to properties, addition of properties, guidelines, documentation in the standard) that would alleviate those issues?

Approach B: TACE-16 repertoire added to Unicode

How would adding the TACE-16 repertoire to Unicode address those issues? And what would be the new problems created by the introduction of that repertoire?

For example:

- dual encoding and stability policy

- does it need to be in the BMP, and if so, how does it fit there?
- would encoding in a non-contiguous area help or hurt compression techniques?

Approach C: TACE-16 repertoire in the PUA

What are the issues that applications are faced with?

For example:

- collisions with other well-established PUA uses, such as CJK:
 - there is not always an "official" mapping, different vendors do different things
 - PUA conflicts:
 - HKSCS 9571 (U+2721B) → U+E78D
 - GB18030 A6D9 (,) → U+E78D
 - PUA differentiation:
 - HKSCS 8BFA (U+20087) → U+F572
 - GB18030 FE51 (U+20087) → U+E816
- PUA characters cannot be used in IDN.

Approach D: TACE-16 as a separate IANA-registered character set

How simple is it to add support for a new character set (with a well-defined mapping to the existing Tamil block) to existing Unicode-based applications? Can this be done in a timely manner, across enough products to achieve viable workflows? What are the implications for already shipped software?

U+0B82 (◌◌) TAMIL SIGN ANUSVARA

U+0B83 (◌◌◌) TAMIL SIGN VISARGA

U+0B85 (அ) TAMIL LETTER A

U+0B86 (ஆ) TAMIL LETTER AA

U+0B87 (இ) TAMIL LETTER I

U+0B88 (ஈ) TAMIL LETTER II

U+0B89 (உ) TAMIL LETTER U

U+0B8A (ஊ) TAMIL LETTER UU

U+0B8E (எ) TAMIL LETTER E

U+0B8F (ஏ) TAMIL LETTER EE

U+0B90 (ஐ) TAMIL LETTER AI

U+0B92 (ஒ) TAMIL LETTER O

U+0B93 (ஓ) TAMIL LETTER OO

U+0B94 (ஔ) TAMIL LETTER AU

U+0B95 (க) TAMIL LETTER KA

U+0B99 (க்ஷ) TAMIL LETTER NGA

U+0B9A (ச) TAMIL LETTER CA

U+0B9C (ஜ) TAMIL LETTER JA

U+0B9E (ஞ) TAMIL LETTER NYA

U+0B9F (ஸ்ரீ) TAMIL LETTER TTA
U+0BA3 (ண) TAMIL LETTER NNA
U+0BA4 (த்ரீ) TAMIL LETTER TA
U+0BA8 (ந்ரீ) TAMIL LETTER NA
U+0BA9 (ண்ரீ) TAMIL LETTER NNNA
U+0BAA (ப்ரீ) TAMIL LETTER PA
U+0BAE (ம்ரீ) TAMIL LETTER MA
U+0BAF (ய்ரீ) TAMIL LETTER YA
U+0BB0 (ர்ரீ) TAMIL LETTER RA
U+0BB1 (ற்ரீ) TAMIL LETTER RRA
U+0BB2 (ல்ரீ) TAMIL LETTER LA
U+0BB3 (ள்ரீ) TAMIL LETTER LLA
U+0BB4 (ழ்ரீ) TAMIL LETTER LLLA
U+0BB5 (வ்ரீ) TAMIL LETTER VA
U+0BB6 (ஶ்ரீ) TAMIL LETTER SHA
U+0BB7 (ஷ்ரீ) TAMIL LETTER SSA
U+0BB8 (ஸ்ரீ) TAMIL LETTER SA
U+0BB9 (ஹ்ரீ) TAMIL LETTER HA
U+0BBE (ா) TAMIL VOWEL SIGN AA
U+0BBF (ி) TAMIL VOWEL SIGN I
U+0BC0 (ீ) TAMIL VOWEL SIGN II
U+0BC1 (ு) TAMIL VOWEL SIGN U
U+0BC2 (ூ) TAMIL VOWEL SIGN UU
U+0BC6 (ெ) TAMIL VOWEL SIGN E
U+0BC7 (ே) TAMIL VOWEL SIGN EE
U+0BC8 (ை) TAMIL VOWEL SIGN AI
U+0BCA (ொ) TAMIL VOWEL SIGN O
U+0BCB (ோ) TAMIL VOWEL SIGN OO
U+0BCC (ௌ) TAMIL VOWEL SIGN AU
U+0BCD (்) TAMIL SIGN VIRAMA
U+0BD7 (ுள்) TAMIL AU LENGTH MARK
U+0BE6 (௦) TAMIL DIGIT ZERO
U+0BE7 (௧) TAMIL DIGIT ONE
U+0BE8 (௨) TAMIL DIGIT TWO
U+0BE9 (௩) TAMIL DIGIT THREE
U+0BEA (௪) TAMIL DIGIT FOUR
U+0BEB (௫) TAMIL DIGIT FIVE
U+0BEC (௬) TAMIL DIGIT SIX
U+0BED (௭) TAMIL DIGIT SEVEN

U+0BEE (௮) TAMIL DIGIT EIGHT
U+0BEF (௯) TAMIL DIGIT NINE
U+0BF0 (௩) TAMIL NUMBER TEN
U+0BF1 (௩௩) TAMIL NUMBER ONE HUNDRED
U+0BF2 (௩௩௩) TAMIL NUMBER ONE THOUSAND
U+0BF3 (௩௩௩௩) TAMIL DAY SIGN
U+0BF4 (௩௩௩௩௩) TAMIL MONTH SIGN
U+0BF5 (௩௩௩௩௩௩) TAMIL YEAR SIGN
U+0BF6 (௩௩௩௩௩௩௩) TAMIL DEBIT SIGN
U+0BF7 (௩௩௩௩௩௩௩௩) TAMIL CREDIT SIGN
U+0BF8 (௩௩௩௩௩௩௩௩௩) TAMIL AS ABOVE SIGN
U+0BF9 (௩௩௩௩௩௩௩௩௩௩) TAMIL RUPEE SIGN
U+0BFA (௩௩௩௩௩௩௩௩௩௩௩) TAMIL NUMBER SIGN