REPORT ON THE

FINAL RECOMMENDATIONS OF THE TASK FORCE ON TACE16

1. OVERVIEW

Initiatives for the usage of Tamil in computers and in Information technology started as early as 1985. But there was no standard encoding for Tamil available then. 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 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.Anandakrishnan 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. During the conference the issues related to the Unicode Tamil were also discussed Based on the recommendations, the Government of Tamil Nadu, in the above G.O., has directed the Sub-Committee on Tamil in Information Technology to propose a 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 Unicode.

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.

1.1 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 aythm. 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. 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 normalization.

- Simple counting, 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 + Ü) 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. சொ is a grapheme; but ச, ெ◌, ◌ா are characters.
- Dynamic Composition - a text element encoded as a sequence of a base character followed by one or more combining marks.
- Unicode Tamil encoding has not followed the Universal principle of Unicode. All the characters of Tamil are not encoded as per the Universal principle of Unicode.
- There are two methods of rendering the following Vowel Consonants.

  è + ª £ = ªè£ = è + ª + £
  è + « £ = «è£ = è + « + £
  è + ª ÷ = ªè÷ = è + ª + ÷

  This leads to ambiguity in rendering characters.

- The Present Unicode is not efficient for sparsing. 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.

- பர்க்கெலி தமிழ்ப் பேராசிரியர் ▯□f◌ார்^ +◌ார்ட், பென்சில்வேனியா பேரா. "◌ிப்மன். ஒரு காலத்தில் ராf◌ாவெல்லாம் முகமூடி போட்டு தனது ராf◌ாங்கத்தின் நிர்வாகத்தை தானே சென்று பார்த்து தவறுகளை உடனுக்குடனேயே செய்து வந்தனர். தமிழர்கள் ஆ?◌ா ஓ□▯?◌ா என்று பேசுவார்கள்.

- This encoding may lead to legal problems. For example, consider the statement "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 ரோf◌ாவுக்குக் கொடுத்து விடுகிறேன்"

  Will the court give the property to திருமதி ரோஜா or திருமதி. ரோஷா?

- The Unicode standard policy is to encode only characters, not glyphs. But, Unicode Tamil standard includes the following vowel signs which are not characters in Tamil language

£, ª, ª £, « £, ª ÷

- Unicode is not supported in many platforms.

1.2. Milestones of the Activities initiated by the TN Government

The above anomalies and problems have made the Tamil Nadu government to seek for changes in the Unicode Tamil. The following actions were taken by the Government to solve the problems in the present Unicode Tamil during the past 8 years:

- The Government of Tamil Nadu announced 8-bit encoding Standards for Tamil (TAM and TAB standards) in June 1999, the Tamil Software development process started. Tamil is the first and the only Indian language to have a unique 8-bit standard among the Indian languages. 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, since these applications were developed in an 8-bit encoding which does not have enough code positions to encode all Tamil characters, they are not as efficient as that of English-based software systems.

- As presented earlier, even the 16-bit encoding for Tamil in Unicode is the same as the 8-bit encoding, and the present Unicode Tamil increases the level of inefficiency and complexity than that caused by the 8-bit encoding. Realizing this limitation of the 8-bit encoding and the present 16-bit Unicode Tamil, the Tamil Nadu 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. (vide G.O.Ms.No.17 dated 13-06-1999). Accordingly, the Tamil Nadu Government initiated action in this direction through Tamil Virtual University (TVU).

- The then Director, TVU formed a committee with experts pooled from KaNithamizh Sangam for this purpose. The committee developed a 16-bit All Character encoding for Tamil and presented the same at the pre-conference session of TamilNet 2000 conference in Colombo, SriLanka 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.

- Meantime the Government of Tamil Nadu initiated action through 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 to ‘Unicode’, the Government of Tamil Nadu renewed its membership as an Associate Member in the Unicode Consortium. The 16-bit character encoding, developed and evaluated through TVU initiative, 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, The IT Secretary, Government of Tamil Nadu Government of India forwarded the proposal to the Government of India for onwards 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.

- Subsequently the MIT, GOI convened a meeting during 19th and 23rd April 2001 to discuss and consolidate the revision of Unicode standard 3.0 incorporated Indian scripts, so as to enable Dr. OmVikas, Director, MIT to present the recommendations at the Technical Committee in the meeting of the Unicode Consortium during 21st to 24th May 2001. Dr. M. Ponnavaikko, Director, Tamil Virtual University participated in the meeting and communicated that the present Unicode standard for Tamil is not adequate for efficient computing and recommended to incorporate the all character encoding for Tamil, recommended by the Tamil Nadu Government.

- Based on the advice of the Unicode consortium communicated by the Ministry of Information Technology, Government of India and on the recommendation of Dr.M. Anandakrishnan, the then, Vice-Chairman, IT Taskforce Committee of the Government of Tamil Nadu and the Vice Chairman, Tamil State Council for Higher Education in his letter dated 25th May 2001, TUV took action for testing the proposed encoding.

- The following two types of schemes were considered as an alternative to the present Unicode Tamil.

  (i) An all Character 16-bit encoding scheme for Tamil.
  (ii) A vowel and consonant scheme.

  To evaluate these schemes the following investigations were made through a professional vendor:

  (i) Space occupation in the Encoding Schemes.
  (ii) Efficiency in Text Processing.
  (iii) Efficiency in Database Application
  (iv) Efficiency in Morphological Analysis.

  The present Unicode Tamil was also subjected to the above tests. The results of the investigation were favorable to the 16-bit All Character Encoding for Tamil (Annexure-1).

- Meantime, MIT GOI finalized the recommendations in the 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, has been arguing 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.Ponnavaikko, 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).

- The IT Secretary, Government of Tamil Nadu wrote on 17th September 2003 to the Director, MIT, GOI, stating that the Government of Tamil Nadu is for incorporating the All Character Encoding for Tamil in the Unicode standard and hence stressed that the Government of India should be firm in its position so that Unicode consortium accept the view of Tamil Diaspora.

- 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).

- As suggested by the Unicode Consortium the proposed all character encoding scheme was placed 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 TUNE were developed in order to test the new encoding in PUA under various operating platforms and in all possible applications. The

TUNE was thus placed in the Unicode block **E200 to E38F** on 24th June 2005. The Tamil computing software personnel from the Tamil Diaspora have effectively tested the scheme and found 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.Ponnavaikko 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, 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". (A copy of the resolution is enclosed for reference).

- 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 the16-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.

2. CONSTITUTION OF A TASK FORCE

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, under the Chairmanship of Dr.M.Anandakrishnan with Dr.M.Ponnavaikko 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 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.

- A member of the task force should be identified for planning the action plan for conducting the tests and for monitoring the progress. Accordingly the test areas and members responsible for the tests are identified as follows:

(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.

- Transparencies in testing should be emphasised 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

Mean time 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.Ponnavaikko and Mr.Manimanivannan along with Pankaj Agrawala, Joint Secretatry, 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-2) 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 though 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 bellow.

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 Old 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. FEASIBILTY 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 (Tolkappiyam, Elu. 17-18). But 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.

3.1. Feasibility as a National Standard

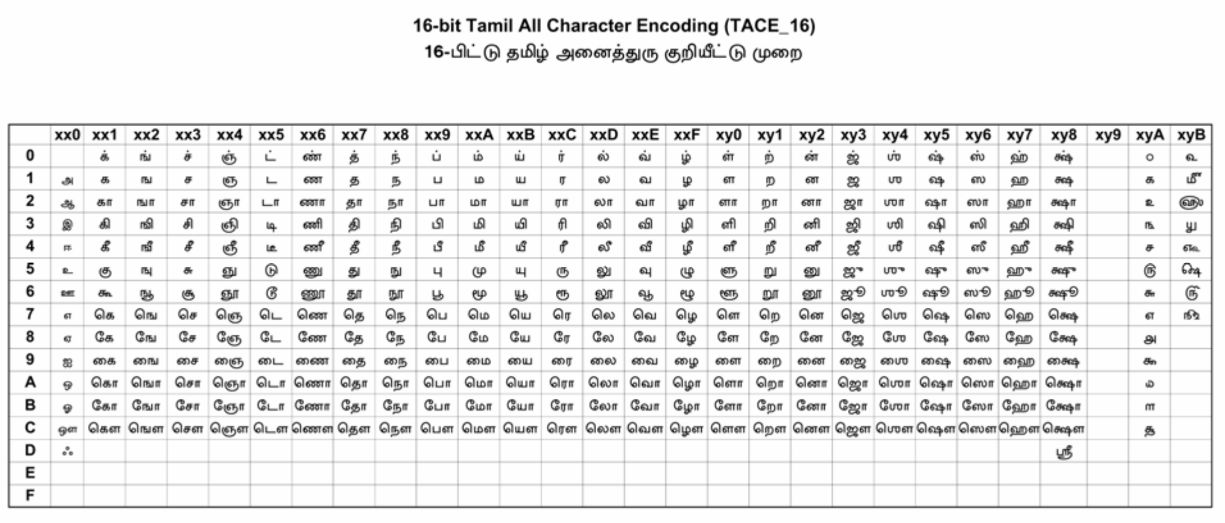
The Tamil Nadu Government declared in 1999 a Bi-lingual Standard TAB and a Mono-lingual Standard TAM for Tamil Language, the only Indian Language to have an 8-bit encoding standard. Over the past 8 years from 1999 till date, a number of Tamil software vendors have developed varieties of application software 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 need to be compensated for converting their software systems into the new 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.

3.2 Feasibility of TACE16 as a Unicode Standard

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

3.2.1 Feasibility of Implementing the old TACE16

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



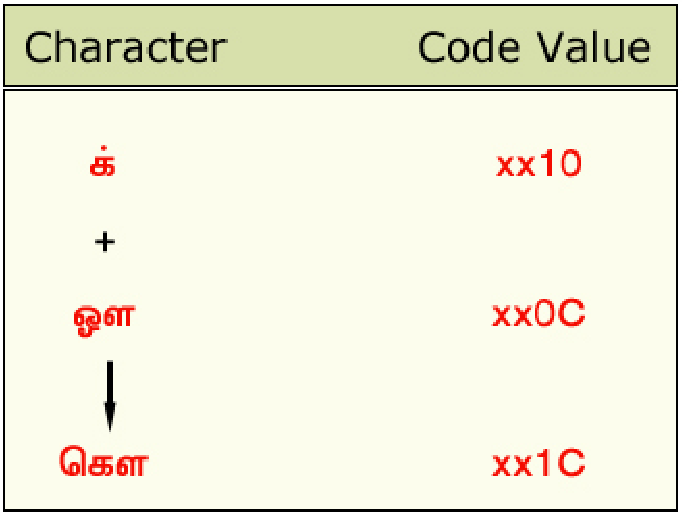
16-bit Tamil All Character Encoding (TACE_16)

16-பிட்டு 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	ண்	த்	ந்	ப்	ம்	ய்	ர்	ல்	வ்	ழ்	ள்	ற்	ன்	ஜ்	ஶ்	ஷ்	ஸ்	ஹ்	க்ஷ்		௦	௳
1	அ	க	ங	ச	ஞ	ட	ண	த	ந	ப	ம	ய	ர	ல	வ	ழ	ள	ற	ன	ஜ	ஶ	ஷ	ஸ	ஹ	க்ஷ		௧	௴
2	ஆ	கா	ஙா	சா	ஞா	டா	ணா	தா	நா	பா	மா	யா	ரா	லா	வா	ழா	ளா	றா	னா	ஜா	ஶா	ஷா	ஸா	ஹா	க்ஷா		௨	௵
3	இ	கி	ஙி	சி	ஞி	டி	ணி	தி	நி	பி	மி	யி	ரி	லி	வி	ழி	ளி	றி	னி	ஜி	ஶி	ஷி	ஸி	ஹி	க்ஷி		௩	௶
4	ஈ	கீ	ஙீ	சீ	ஞீ	டீ	ணீ	தீ	நீ	பீ	மீ	யீ	ரீ	லீ	வீ	ழீ	ளீ	றீ	னீ	ஜீ	ஶீ	ஷீ	ஸீ	ஹீ	க்ஷீ		௪	௷
5	உ	கு	ஙு	சு	ஞு	டு	ணு	து	நு	பு	மு	யு	ரு	லு	வு	ழு	ளு	று	னு	ஜு	ஶு	ஷு	ஸு	ஹு	க்ஷு		௫	௸
6	ஊ	கூ	ஙூ	சூ	ஞூ	டூ	ணூ	தூ	நூ	பூ	மூ	யூ	ரூ	லூ	வூ	ழூ	ளூ	றூ	னூ	ஜூ	ஶூ	ஷூ	ஸூ	ஹூ	க்ஷூ		௬	௹
7	எ	கெ	ஙெ	செ	ஞெ	டெ	ணெ	தெ	நெ	பெ	மெ	யெ	ரெ	லெ	வெ	ழெ	ளெ	றெ	னெ	ஜெ	ஶெ	ஷெ	ஸெ	ஹெ	க்ஷெ		௭	௺
8	ஏ	கே	ஙே	சே	ஞே	டே	ணே	தே	நே	பே	மே	யே	ரே	லே	வே	ழே	ளே	றே	னே	ஜே	ஶே	ஷே	ஸே	ஹே	க்ஷே		௮	
9	ஐ	கை	ஙை	சை	ஞை	டை	ணை	தை	நை	பை	மை	யை	ரை	லை	வை	ழை	ளை	றை	னை	ஜை	ஶை	ஷை	ஸை	ஹை	க்ஷை		௯	
A	ஒ	கொ	ஙொ	சொ	ஞொ	டொ	ணொ	தொ	நொ	பொ	மொ	யொ	ரொ	லொ	வொ	ழொ	ளொ	றொ	னொ	ஜொ	ஶொ	ஷொ	ஸொ	ஹொ	க்ஷொ		௰	
B	ஓ	கோ	ஙோ	சோ	ஞோ	டோ	ணோ	தோ	நோ	போ	மோ	யோ	ரோ	லோ	வோ	ழோ	ளோ	றோ	னோ	ஜோ	ஶோ	ஷோ	ஸோ	ஹோ	க்ஷோ		௱	
C	ஔ	கௌ	ஙௌ	சௌ	ஞௌ	டௌ	ணௌ	தௌ	நௌ	பௌ	மௌ	யௌ	ரௌ	லௌ	வௌ	ழௌ	ளௌ	றௌ	னௌ	ஜௌ	ஶௌ	ஷௌ	ஸௌ	ஹௌ	க்ஷௌ		௲	
D	ஃ																								ஸ்ரீ			
E																												
F																												

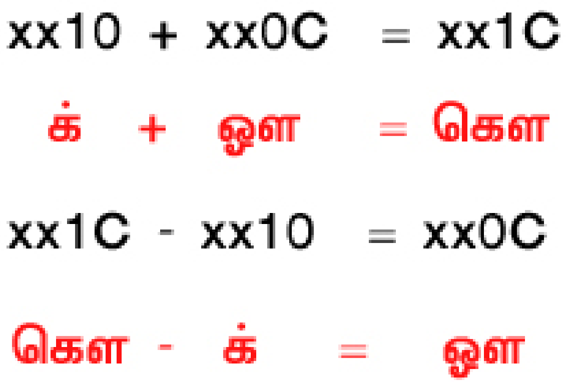
This system has the following advantages:

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



Character	Code Value
க்	xx10
+	
ஔ	xx0C
↓	
கௌ	xx1C

- ✓ By simple arithmetic operation the characters can be parsed



xx10 + xx0C = xx1C

க் + ஔ = கௌ

xx1C - xx10 = xx0C

கௌ - க் = ஔ

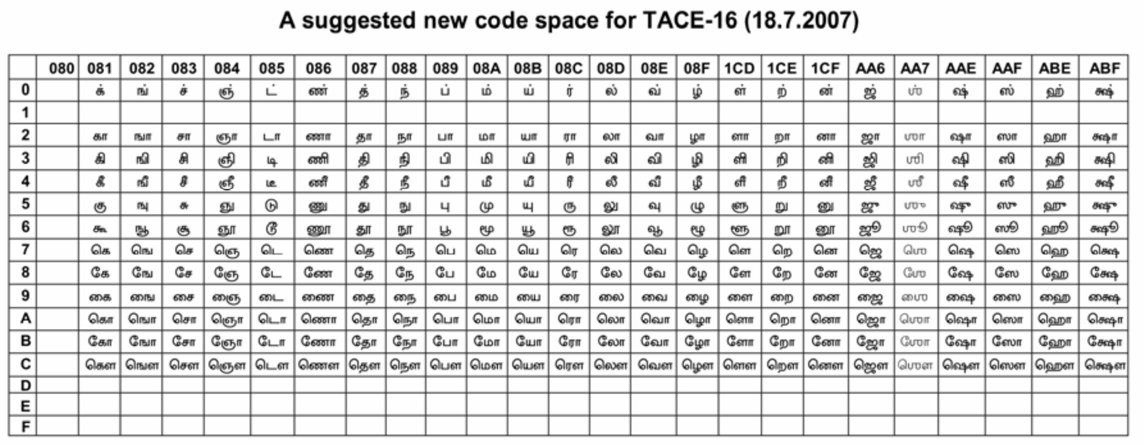
- ✓ Sorting and searching is very simple.
- ✓ The Collation is sequential in accordance with the code value.
- ✓ The encoding is unambiguous.
- ✓ Any given code point always represents the same character.
- ✓ There is no ambiguity as in the Present Unicode Tamil.

But there are feasibility issues in implementing this scheme in the BMP area of the Unicode space. They 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.

3.2.2. 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)

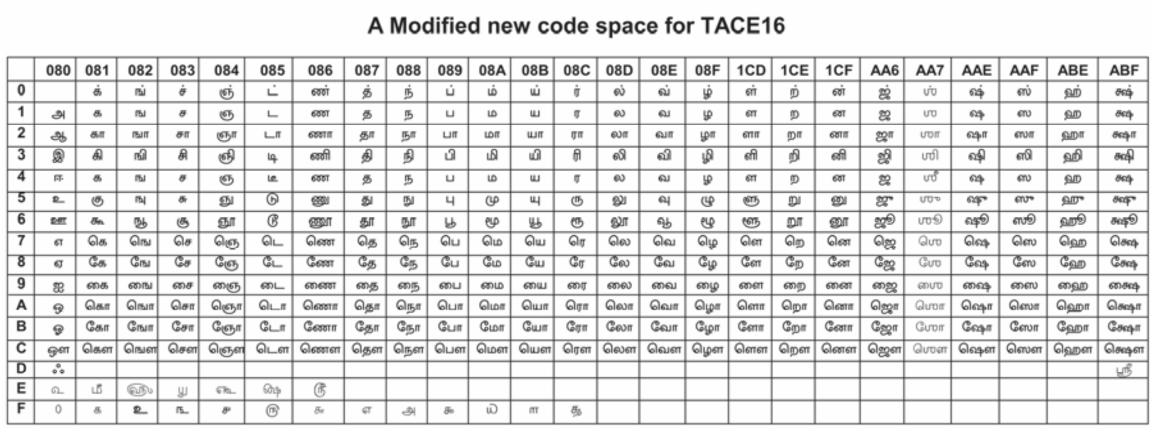
	080	081	082	083	084	085	086	087	088	089	08A	08B	08C	08D	08E	08F	1CD	1CE	1CF	AA6	AA7	AAE	AAF	ABE	ABF
0		க்	ங்	ச்	ஞ்	ட்	ண்	த்	ந்	ப்	ம்	ய்	ர்	ல்	வ்	ழ்	ள்	ற்	ன்	ஜ்	ஶ்	ஷ்	ஸ்	ஹ்	க்ஷ்
1																									
2		கா	ஙா	சா	ஞா	டா	ணா	தா	நா	பா	மா	யா	ரா	லா	வா	ழா	ளா	றா	னா	ஜா	ஶா	ஷா	ஸா	ஹா	க்ஷா
3		கி	ஙி	சி	ஞி	டி	ணி	தி	நி	பி	மி	யி	ரி	லி	வி	ழி	ளி	றி	னி	ஜி	ஶி	ஷி	ஸி	ஹி	க்ஷி
4		கீ	ஙீ	சீ	ஞீ	டீ	ணீ	தீ	நீ	பீ	மீ	யீ	ரீ	லீ	வீ	ழீ	ளீ	றீ	னீ	ஜீ	ஶீ	ஷீ	ஸீ	ஹீ	க்ஷீ
5		கு	ஙு	சு	ஞு	டு	ணு	து	நு	பு	மு	யு	ரு	லு	வு	ழு	ளு	று	னு	ஜு	ஶு	ஷு	ஸு	ஹு	க்ஷு
6		கூ	ஙூ	சூ	ஞூ	டூ	ணூ	தூ	நூ	பூ	மூ	யூ	ரூ	லூ	வூ	ழூ	ளூ	றூ	னூ	ஜூ	ஶூ	ஷூ	ஸூ	ஹூ	க்ஷூ
7		கெ	ஙெ	செ	ஞெ	டெ	ணெ	தெ	நெ	பெ	மெ	யெ	ரெ	லெ	வெ	ழெ	ளெ	றெ	னெ	ஜெ	ஶெ	ஷெ	ஸெ	ஹெ	க்ஷெ
8		கே	ஙே	சே	ஞே	டே	ணே	தே	நே	பே	மே	யே	ரே	லே	வே	ழே	ளே	றே	னே	ஜே	ஶே	ஷே	ஸே	ஹே	க்ஷே
9		கை	ஙை	சை	ஞை	டை	ணை	தை	நை	பை	மை	யை	ரை	லை	வை	ழை	ளை	றை	னை	ஜை	ஶை	ஷை	ஸை	ஹை	க்ஷை
A		கொ	ஙொ	சொ	ஞொ	டொ	ணொ	தொ	நொ	பொ	மொ	யொ	ரொ	லொ	வொ	ழொ	ளொ	றொ	னொ	ஜொ	ஶொ	ஷொ	ஸொ	ஹொ	க்ஷொ
B		கோ	ஙோ	சோ	ஞோ	டோ	ணோ	தோ	நோ	போ	மோ	யோ	ரோ	லோ	வோ	ழோ	ளோ	றோ	னோ	ஜோ	ஶோ	ஷோ	ஸோ	ஹோ	க்ஷோ
C		கௌ	ஙௌ	சௌ	ஞௌ	டௌ	ணௌ	தௌ	நௌ	பௌ	மௌ	யௌ	ரௌ	லௌ	வௌ	ழௌ	ளௌ	றௌ	னௌ	ஜௌ	ஶௌ	ஷௌ	ஸௌ	ஹௌ	க்ஷௌ
D																									
E																									
F																									

The following are the feasibility 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, AAFF, 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)

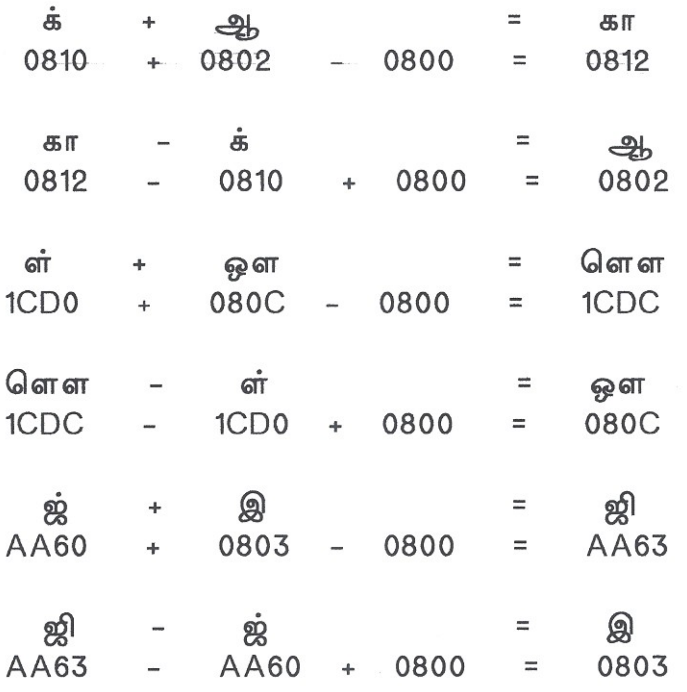


A Modified new code space for TACE16

	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	ஶ்	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	ஶ	ஷ	ஸ	ஹ	க்ஷ
2	ஆ	கா	ஙா	சா	ஞா	டா	ணா	தா	நா	பா	மா	யா	ரா	லா	வா	ழா	ளா	றா	னா	ஜா	ஶா	ஷா	ஸா	ஹா	க்ஷா
3	இ	கி	ஙி	சி	ஞி	டி	ணி	தி	நி	பி	மி	யி	ரி	லி	வி	ழி	ளி	றி	னி	ஜி	ஶி	ஷி	ஸி	ஹி	க்ஷி
4	ஈ	க	ங	ச	ஞ	டீ	ண	த	ந	ப	ம	ய	ர	ல	வ	ழ	ள	ற	ன	ஜ	ஶீ	ஷ	ஸ	ஹ	க்ஷ
5	உ	கு	ஙு	சு	ஞு	டு	ணு	து	நு	பு	மு	யு	ரு	லு	வு	ழு	ளு	று	னு	ஜு	ஶு	ஷு	ஸு	ஹு	க்ஷு
6	ஊ	கூ	ஙூ	சூ	ஞூ	டூ	ணூ	தூ	நூ	பூ	மூ	யூ	ரூ	லூ	வூ	ழூ	ளூ	றூ	னூ	ஜூ	ஶூ	ஷூ	ஸூ	ஹூ	க்ஷூ
7	எ	கெ	ஙெ	செ	ஞெ	டெ	ணெ	தெ	நெ	பெ	மெ	யெ	ரெ	லெ	வெ	ழெ	ளெ	றெ	னெ	ஜெ	ஶெ	ஷெ	ஸெ	ஹெ	க்ஷெ
8	ஏ	கே	ஙே	சே	ஞே	டே	ணே	தே	நே	பே	மே	யே	ரே	லே	வே	ழே	ளே	றே	னே	ஜே	ஶே	ஷே	ஸே	ஹே	க்ஷே
9	ஐ	கை	ஙை	சை	ஞை	டை	ணை	தை	நை	பை	மை	யை	ரை	லை	வை	ழை	ளை	றை	னை	ஜை	ஶை	ஷை	ஸை	ஹை	க்ஷை
A	ஒ	கொ	ஙொ	சொ	ஞொ	டொ	ணொ	தொ	நொ	பொ	மொ	யொ	ரொ	லொ	வொ	ழொ	ளொ	றொ	னொ	ஜொ	ஶொ	ஷொ	ஸொ	ஹொ	க்ஷொ
B	ஓ	கோ	ஙோ	சோ	ஞோ	டோ	ணோ	தோ	நோ	போ	மோ	யோ	ரோ	லோ	வோ	ழோ	ளோ	றோ	னோ	ஜோ	ஶோ	ஷோ	ஸோ	ஹோ	க்ஷோ
C	ஔ	கௌ	ஙௌ	சௌ	ஞௌ	டௌ	ணௌ	தௌ	நௌ	பௌ	மௌ	யௌ	ரௌ	லௌ	வௌ	ழௌ	ளௌ	றௌ	னௌ	ஜௌ	ஶௌ	ஷௌ	ஸௌ	ஹௌ	க்ஷௌ
D	ஃ																								ஸ்ரீ
E	௳	௴	௵	௶	௷	௸	௹																		
F	௦	௧	௨	௩	௪	௫	௬	௭	௮	௯	௰	௱	௲												

- 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:

–



க் + ஆ = கா
0810 + 0802 – 0800 = 0812

கா – க் = ஆ
0812 – 0810 + 0800 = 0802

ள் + ஔ = ளௌ
1CD0 + 080C – 0800 = 1CDC

ளௌ – ள் = ஔ
1CDC – 1CD0 + 0800 = 080C

ஜ் + இ = ஜி
AA60 + 0803 – 0800 = AA63

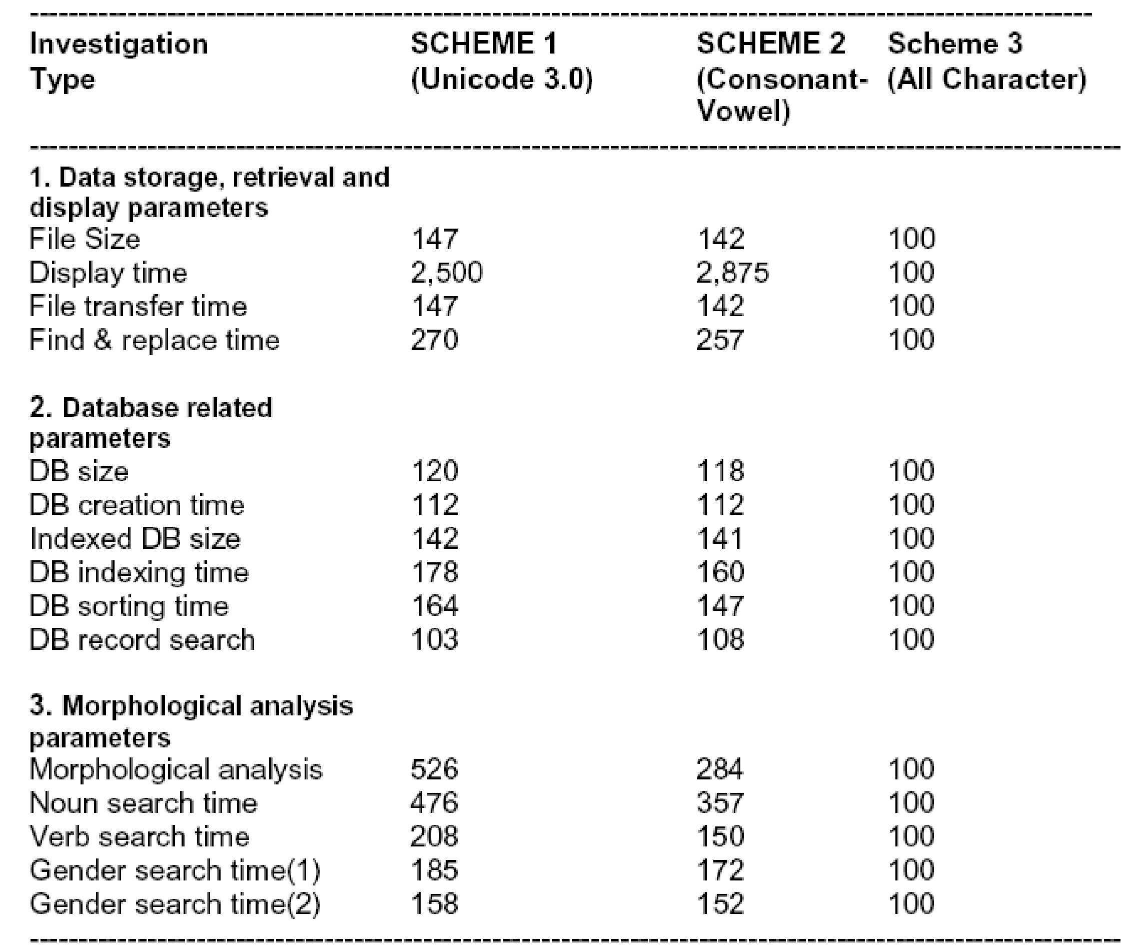
ஜி – ஜ் = இ
AA63 – AA60 + 0800 = 0803

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.

4. 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.

Annexure-1



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

Annexure-2

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. Ponnavaikko (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.