Unicode Tamil - Evolution of Alternate 16 bit encoding scheme to solve the problems in the current scheme

A. Elangovan,
Cadgraf Digitals Private Limited; Elango Tamil software; Chennai
Chair, INFITT Working Group (WG08) for Alternate 16 bit encoding scheme for Tamil
Dr. V. Sankaranarayanan,
Director, Tamil Virtual University, Chennai

Abstract:

The Unicode encoding for Tamil, in spite of being aggressively pushed by Microsoft, is yet to become widely used by one and all. The reasons for this slow progress are many. Unlike other Indian languages Tamil has no Complex scripts. But Tamil script has been defined as Complex script in current Unicode. This has resulted in Tamil requiring Level-2 implementation instead of Level-1 implementation needed for non-complex scripts. Tamil Unicode support is yet to be available in other operating systems like MacOS, Linux, etc; Major Application developers like Adobe are yet to provide Tamil Unicode support even in Windows; The collation order for Tamil is incorrectly implemented even in MS Office: Unicode fonts developed for one application(eg.OfficeXP) will not work in other Unicode aware applications(eg. Adobe InDesign).

The Alternate 16 bit encoding has been developed primarily to solve these problems. It is based on the fact that Tamil script is a Simple script and not a Complex script. "Unicode Tamil-New", a full-fledged Keyboard Interface and a Unicode font has been developed based on the "Alternate 16 bit Encoding Scheme". The initial trials indicate that most of the above mentioned problems of current Tamil Unicode are solved. The Unicode Tamil-New encoding works satisfactorily in most of the applications like OpenOffice, all browsers, Adobe, Quark, Corel, Macromedia apart from Microsoft applications. The encoding is also compatible to Windows98, 2000, NT, XP, ME, MacOS9, MacOSX versions. It is expected to be suitable for Unix and Linux also. Natural and correct sorting order is provided in the encoding. The fonts are very easy to develop and easy to implement in all applications without waiting for the respective developers to provide for Tamil support. It is about 40% more efficient than the current Standard Unicode Tamil. It is found to be the cheapest way to provide Tamil IT enabled services and the fastest way to spread Tamil usage in all IT and communication software.

1. Introduction to Font Encoding

A computer stores text in the form of numbers and not in the visual written form. Each alphabet is stored internally as a number, but displayed on the screen as a **Glyph** (written form). A **Font** is used to convert the internal number to a visual form on the screen. By changing the font one can use the text in any type face. By changing the shapes defined in the font, the computer can represent any language. For a long time a 'BYTE' (consisting of 8 bits) was the basic unit of a character in the computer. A BYTE could accommodate 256

different characters. Out of these 256 locations a minimum of 32 locations are reserved for use by the operating system. Hence only 224 locations are freely available. To represent a language in the computer one had to allot each one of these locations to a character or glyph of the language. This process of allotting a location to a character or glyph is called **Encoding**.

The current English language encoding in the Windows operating system is the ANSI scheme. Since the English alphabets contains only 52 letters (26 capital letters and 26 small letters), the assignment of numbers to locations is straight forward. The Tamil language has 313 alphabets including the Grantha alphabets. All though it is desirable to allot one location for each alphabet, we cannot do so due to limitations of space (224 locations). Hence we assign one location for each glyph (e.g. kaal, kombu, kokki, pulli etc.). This system is called the "**Glyph Encoding**". By doing this we reduce the number of locations required to implement the Tamil script. The Government of Tamilnadu has announced two such encoding schemes for Tamil. They are the TAM and TAB standards. TAM is a monolingual encoding scheme and the TAB is a bilingual encoding scheme. In these encoding schemes since there is a one to one relation between storage and display (i.e. every byte stored is displayed) the Tamil script could be implemented in any 'readymade' software package without any additional support. While it was relatively easy for the Tamil script to be encoded, because of the lesser number of characters, uniformity of script and absence of 'conjunct consonants', it was more difficult to encode the other Indian languages.

2. ISCII (Indian Script Code for Information Interchange)

The Indian government, with a view to have a common encoding for all Indian languages, developed the ISCII standard.

Features

The features of this encoding scheme are:

- This encoding is bilingual in nature.
- The first 128 locations is exactly as per the ANSI encoding standard and contains the English alphabets, commonly used punctuation marks and symbols.
- The next 128 locations contains the encoding for each Indian language.
- Since these 128 locations are not enough to accommodate the alphabets of most of the Indian languages, only vowels, consonants and vowel modifiers were encoded.
- Similar vowels, consonants and vowel modifiers of various languages occupied the same slot. i.e the vowel 'a' would be allocated the same location in all the languages.
- This system of encoding enabled viewing of the text in various of languages just by change of the font. A text typed in Tamil could be easily read in a transliterated form in Hindi by using a Hindi font. It may not be true vice-versa since all the Hindi consonants do not have a Tamil equivalent.

Disadvantages

The major disadvantages of this system are:

- It requires more space since almost all the alphabets would be stored in their broken down form.
- It does not have a one-to-one relation between the stored bytes and displayed bytes. For example the Tamil alphabet 'ku' would be displayed as a single glyph, but it would be stored internally as a consonant 'k' + vowel modifier 'u'. There is a two-to-one relation between the stored text and the display.
- This added complexity in display makes it unsuitable for usage in 'ready-made' software.
- Thus it was not used in Tamil.

It must be noted that the same 256 locations are used by different languages. Hence it is apparent that unless we know which language the text pertains to, we cannot use the appropriate font to view it. This led to a chaotic situation. In order to avoid this confusion, alphabets of different languages had to be given different numbers. This required more locations. Thus was born the 'double byte' encoding scheme which uses 16 bits to represent a character instead of 8 bits. In the 16 bit space 65,536 locations are available as compared to 256 locations in the 8 bit space.

3. Unicode

Unicode is a 16 bit encoding scheme which is the most common 16 bit encoding scheme in use today. It contains characters of all the major world languages. It is being developed by the Unicode Consortium which has the major software developers and computer manufacturers as members. The Indian Government and TN Government are members in the consortium. It is a stated policy of Unicode that only characters will be encoded and not glyphs. It also states that it is not concerned with efficiency of the encoding system. It must be noted however that in the beginning all existing standard encoding schemes of different languages were implemented 'as is' in Unicode without consideration of the above principles. Because of this policy the ISCII standard which was primarily designed for an 8-bit environment was used as the base for implementing the Indian languages in Unicode. Thus the Tamil block of Unicode was based on the Tamil encoding in ISCII which was not used at all till then. A Simple script was converted into a Complex script.

4. Simple and Complex Scripts

In some Indian languages, when two consonants come together, like ik and ir, the glyphs are not rendered in the normal way, but rendered in a different way. In these languages, for one character, more than one glyph rendering is possible. It depends on the situation. These languages are said to have **Complex Scripts**. For these languages, the character type representation in memory may be beneficial. But they have to pay the price of a rendering module. In Tamil for each character, there is only one way of rendering. **Hence Tamil does NOT have a Complex Script.**

Because of the necessity of a processing module to show the letters on the screen and coordinate with the memory content, any software designed for English will not work as such for complex scripts. But they will work smoothly for simple scripts. This is the reason for glyph encoding being popular in India, whereas ISCII has not become that much popular in the commercial world. **In Tamil the use of ISCII is negligible.**

5. The Unicode (16-bit) Encoding Scheme for Tamil

65,536 combinations are possible with 16 bits. If 16 bits are used for every character, then many languages of the world can be accommodated in this scheme. Unicode is designed to accommodate many languages of the world**. It is supposed to encode characters and not glyphs.**

5.1 Formation of Indic blocks from ISCII

As stated above, in ISCII each Indian language was given 128 slots. Basically the same scheme was adopted in Unicode also. Each block is different from the other block. Hence the perceived advantage of ISCII -'store in one language and see in any language' is not valid in Unicode. **But the disadvantages of ISCII have been carried over with further setback.** For Indian languages having complex scripts, there may not be much difference in using the Unicode instead of ISCII. **But Tamil does NOT have a complex script.**

In ISCII, like others, Tamil is also given 128 slots. Tamil having a simple script, should have been treated differently in the Unicode. All its 313 characters and the special symbols could have been accommodated easily. But unfortunately, in Unicode, Tamil is given 128 slots only. This has resulted in many disadvantages.

6. Disadvantages of current Unicode

6.1. Not a real character encoding

As already mentioned, this coding taken from ISCII is not character encoding. Unicode is supposed to encode characters, but this is not the case. Pure consonants, which are fundamental in nature, do not have single slots. Pure consonants have been treated in an unnatural way. This will be a constant irritant to the programmers while doing natural language processing in a large scale, in the future.

6.2. Multiple ways in representation

For some letters like ko, there seems to be two different ways of coding. One as ka and the vowel modifier for o. The other is ka plus the vowel modifier for ae plus the vowel modifier for aa. This ambiguity in representation will lead to a situation where search and similar operations can lead to incorrect results.

6.3. Not a complex script

Above all, treating Tamil as having a complex script has resulted in Level2 implementation in Unicode with enormous negative consequences, which will hinder the growth of the language use in the future.

6.4 Dependence on OS and Application developers for Unicode Tamil

The complex scripts which are implemented in level2 in Unicode require the support of developers of OS and Applications for Tamil Unicode at the core level. While Level-1 support for simple scripts is available in almost all current Unicode aware OS and Applications, Level-2 support for complex scripts is not available in most of them except a few like MS OfficeXP, OpenOffice,etc:. This puts a permanent dependence on OS and Application developers for Tamil support in Unicode.

It is over 14 years since Unicode Tamil has been announced as standard. Adobe (one of the founder member of Unicode consortium), the leading developer of the publishing software in the world, when contacted, has stated categorically that they have no plans to provide Indic support in the near future. Though over 100 Unicode Tamil Fonts are developed by third party Tamil software developers, the users of many popular applications like Indesign, PageMaker, Photoshop,etc; have NO option but to continue with the 8 bit encoding. Unicode Tamil is a long wait for them. Unless sufficient business volume in Tamil is expected, it may take decades for these developers to provide the Tamil support.

7. Evolution of Alternate 16 bit Encoding for Tamil

7.1 Tamil Virtual University

The inadequacies of the current Unicode for Tamil was brought out in various forums. After a detailed study by Tamil Virtual University, a new code block containing 384 codes with all character representation for Tamil was evolved. Various types of tests were conducted for text, database and linguistic applications. The test results showed a very clear advantage for the new scheme in terms of processing speed, storage space and transmission speed.

7.2 TN Govt, MIT and Unicode Consortium

The TN Government had submitted the above scheme to the MIT (Ministry of Information Technology, New Delhi) and the Unicode Consortium for consideration. Unicode Consortium citing the reasons of code stability in-spite of its advantages did not accept the code. However the Unicode consortium suggested an alternate measure to put the proposed All Character scheme in Private Use Area (PUA). As a follow up of this, the All Character 16 bit encoding for Tamil was fine tuned for putting on the Unicode Private Use Area. A new keyboard driver and a new Unicode font were developed in order to test the new encoding in PUA in various Operating Systems and Applications.

7.3 INFITT Working Group - WG08:

The INFITT EC has formed a separate Working Group (WG08) on 1st January 2004 to evaluate the "Alternate 16 bit encoding scheme for Tamil" proposed by TVU. A new Unicode Tamil font making use of the the PUA and a new keyboard driver for the Unicode Tamil-New are developed and tests carried out in different applications and operating systems. The initial trials show very encouraging results and the testing is reaching the stage of beta very shortly wherein the the keyboard and the font for the new scheme will be made available in the Public Domain for testing by others. After getting the feedback from all beta testers, it will converted into a formal RFC.

7.4 Code Block for "Tamil Unicode-New" in the Private Use Area:

- The New Unicode Tamil block proposed for incorporating in the Private Use Area, viz Tamil Unicode New is furnished in annexure-1A & 1B.
- The earmarked PUA starts from E000 and extends up to E8FF. Since, some of the early occupiers of the PUA might have started from the initial blocks, it is proposed to leave the early blocks and use the block E200 to E38F for the Tamil Unicode New.
- Tamil Unicode -New in PUA will co-exist with the Tamil Unicode Standard ver 4.0 in Unicode regular space. Hence, all Tamil characters including Grandha letter only are included in the new block as in Annexure-1; for all other special symbols, the existing locations in the current Unicode standard ver 4.0 will be made use of.
- In the character set, one more character "SHA" has been added. This character and Tamil numeral for Zero are already in the pipeline table announced by Unicode Standard 4.0.
- The order and arrangement of the Grandha characters are as per the sorting order recommended to the Tamilnadu Government recently by the Tamil Virtual University.

8. Advantages of the Tamil Unicode-New

8.1. Simple Script and Level 1 Implementation

In Unicode Tamil-New, the Tamil script is treated as a Simple Script (non complex) and hence it is implemented as Unicode Level-1 instead of Level 2 implementation in the current Unicode standard.

8.2 No Dependence on OS and Application developers

With Unicode Level-1 implementation, it is not necessary for the OS and Application developers to provide Indic support for Tamil Unicode at the core level. The "TAU_Elango_Bharathi.ttf" Unicode new font for Tamil is found to be working in most of the Operating systems like Windows 2000/NT/XP, Windows 98 and Mac OS09/ OSX. The Unix and Linux are under testing. Also it is found to be working well with most of the popular packages like MS Office, Open Office, Adobe In Design, Quark Xpress, Photoshop, Illustrator, Outlook Express, Internet Explorer, etc; It is expected that any current or future

Unicode aware applications will be able to support Unicode Tamil-New automatically. Dependency on the application developers for Tamil support has been eliminated.

8.3 Natural Sorting Order

All Tamil Character blocks are arranged in the correct sorting order as per the proposal to the TN Govt. Hence all applications will give correct and uniform sorting order without any algorithm to be provided by the application provider.

8.4 Real Character Encoding

It is the real character encoding, representing the true nature of Tamil as they appear in primary school Tamil text books.

8.2. Efficient Design

The creation of the 16 bits is done in a scientific way. Of the sixteen bits, the first 7 bits indicates the language. The next 5 bits gives the serial number of the consonant part of a Tamil letter. The next 4 bits gives the serial number of the vowel part of a Tamil letter. A zero here means the absence of the consonant or vowel part, that is, it is a pure vowel or pure consonant. Hence, it is extremely easy to see what a letter contains. This simplicity comes from the natural way in which the coding is designed. Unicode Tamil-New is as efficient as the English in all applications tested so far. It is at least 40% more efficient than the current Unicode Standard ver 4.0. Unicode Consortium may not be concerned about efficiency of encoding. Can Tamil Diaspora afford to pass this avoidable inefficiency to the next generations?

8.3. Savings in Cost of Computer Storage Space

Its simplicity leads to enormous savings. The space requirement for a Unicode Tamil-New is about 40% less than what is required in the current Unicode Tamil. If we calculate the cumulative storage requirement of over 60 million Tamils, it will be in thousands of crores of rupees every month.

8.4. Saving in Cost of Internet Communication Bandwidth

The Time and Cost required to communicate Tamil text in the Unicode Tamil-New encoding is about 40% less than in the current Unicode.

8.5. Saving in Computer Display Time

The display time required to display the stored data from the hard disk to the display monitor is less than one tenth of that in the current Unicode scheme. The extra time of waiting for display could be very high if the file size is larger.

8.6. Other Savings

The rendering time, searching time and many language processing times are also significantly less in the Unicode Tamil-New as compared to current Unicode.

8.7. Lesser Overall Cost

Unicode Tamil-New will cost much less in both initial cost of hardware and software as well as in the recurring cost of providing the IT enabled services to the masses. It may be argued by some that with the falling prices of hardware and abundant bandwidth availability, cost is not a criterion anymore. The comparative cost will always be at-least 40% less compared to the current Unicode. This is an achievable, recurring saving in expenditure. Current Unicode will result in enormous drain of the resources of the Tamil community. Thousands of crores of rupees will be wasted every month forever until we take corrective action.

Conclusion

Tamil Diaspora in developing countries are interested in the fastest spread of Tamil in all forms of IT enabled services at the lowest cost possible. Cost of providing IT services is a primary concern in many developing countries including India especially when we want to reach to the masses in the rural areas.

The Tamil Unicode-New which takes into account all aspects of requirements for incorporating in the Unicode Private Use Area is found to be fully adequate and strongly recommended for implementation in the proposed Unicode Private Use Area E200-E38F.

Acknowledgement

The authors acknowledge with thanks Dr.V. Krishnamoorthy, Department of Computer Science, Cresent Engineering College, Chennai, Thiru. P. Chellappan, Palaniappa Brothers, Chennai, Thiru. P. Srinivasan, Senior Consultant, Tamil Virtual University, Chennai for their valuable contributions.

Annexture - 1 - A
TAU_Elango_Barathi New
UNICODE TAMIL NEW ENCODING
(TAMIL CHARACTERS BLOCK IN PRIVATE USE AREA)

	E20	E21	E22	E23	E24	E25	E26	E27	E28	E29	E2A	E2B	E2C	E2D	E2E	E2F	E30	E31	E32	E33	E34	E35	E36	E37	E38
0	Null																								
1																									
2																									
3																									
4																									
5																									
6																									
7																									
8																									
9																									
A																									
B																									
C																									
D																									
E																									
F																									

Annexture - 1 - B
TAU_Elango_Barathi New
UNICODE TAMIL NEW ENCODING
(TAMIL SYMBOLS BLOCK IN EXISTING AREA)

	0B8	0B9	0BA	0BB	0BC	0BD	0BE	0BF
0					◌			○
1					◌			◌
2	◌				◌			◌
3	◌							◌
4								◌
5								◌
6					◌		◌	◌
7					◌		◌	◌
8					◌		◌	◌
9							◌	◌
A					◌		◌	◌
B					◌		◌	
C					◌		◌	
D					◌		◌	
E				◌			◌	
F				◌			◌	