Standardization of Encoding Conversion

P.Chellappan < chellappan@vsnl.com>
Palaniappa Bros, Chennai

Synopsis

Finally a stage has been reached where there are four recognized encoding schemes for Tamil. They are TAM, TAB, TSCII and UNICODE. In addition there are also a number of other propriety encoding schemes still in practice. It will take some more time before the other schemes are disbanded in favor of the recognized schemes.
In this scenario conversion of data from one scheme to another becomes imperative. While conversion of the basic Tamil text is relatively simple, problems arise when certain numerals / symbols / characters (herein after referred to as symbols) have to be converted. This is due to the fact that not all these symbols are found in all the encoding schemes. Hence there could be a potential loss of data during round-trip conversion of data from one scheme to another.
Many developers of 'Encoding Converters' have either ignored this issue or have sought to solve the problem in their own way.
This paper seeks to propose a standard way of dealing with this issue, so that developers of converters will encode the missing characters / symbols in a similar way and thus prevent loss of data by making it developer independent.

Missing Symbols in English

The problem of missing symbols is not unique to Tamil. It also exists in English. For example characters 195, 197, 198, 215, 218, 222, 223, 240, 245, 249, 250, 251, 253, 254 and 255 of the Mac Roman encoding are not encoded in the Windows ANSI encoding. Similarly many symbols found in the ANSI character set are not encoded in the Mac Roman character set.

Missing Symbols in Tamil

As already indicated the problem of missing symbols exists in Tamil also. Table-1 lists all the symbols that are not available in one encoding scheme or the other. With so many symbols already not being encoded in one scheme or the other and with the possibility of more symbols being added to the Tamil block of Unicode which are not encoded in the current 8-bit schemes, one can realize the importance of dealing with this issue in a systematic and regulated manner.

Conventional Solution

The most intuitive and straightforward solution to this problem is to either replace these missing symbols with '?' or replace these symbols with their literal equivalents. For example when a Mac Roman text is re-encoded to the Win encoding the ligature 'fi' will be replaced by the two separate characters fi.

A similar approach is being followed in Microsoft Word, when storing ANSI encoded text as plain text. For example character 153 (™) is converted to (TM), character 169 (©) is converted to (C) and character 174 (®) is converted to (R).

The above approach to solve this issue is very simple but obviously there is loss of data and it clearly shows up when one has to re-convert the text to the original encoding. During this process one does not know for sure if the text (TM) existed in the original data also as (TM) or as the symbol (™).

Proposed Solution

The proposed solution to the problem will entirely eliminate any loss of data during round-trip conversion of data from any encoding to another.

In order to achieve this we need to not only have a set of defined standard literal equivalents for each of these symbols but also define one code point in each encoding scheme to denote the start and end of a literal string.

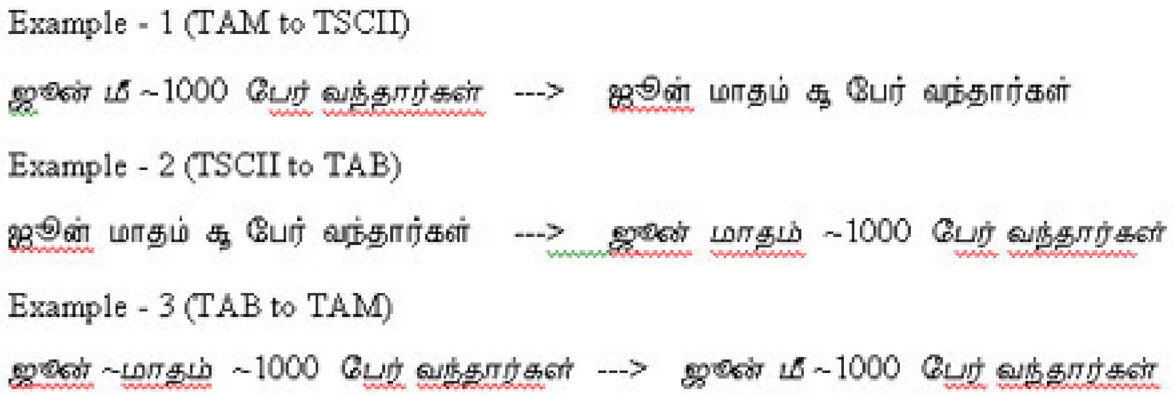
The following code points can be used as a "Literal Start-End Toggle" character in each of the four recognized Tamil encoding schemes:

TAM	:	128
TAB	:	128
TSCII	:	255
UNICODE	:	U+0B80

This special character can be defined as a "Zero Width Blank" character in the font if we want it to be invisible or as any other unique shape that we may desire. For the examples in this paper I have chosen "~" as the glyph for this special character

Table-2 shows all the missing Tamil symbols and their proposed literal equivalents.

Once we have defined the above, the matter becomes very simple. While converting text from one encoding to another, we can replace these missing symbols with their literal equivalents. The following examples illustrates the round-trip conversion process:



Example - 1 (TAM to TSCII)

ஜூன் ௴ ~1000 பேர் வந்தார்கள் ---> ஜூன் மாதம் ௲ பேர் வந்தார்கள்

Example - 2 (TSCII to TAB)

ஜூன் மாதம் ௲ பேர் வந்தார்கள் ---> ஜூன் மாதம் ~1000 பேர் வந்தார்கள்

Example - 3 (TAB to TAM)

ஜூன் ~மாதம் ~1000 பேர் வந்தார்கள் ---> ஜூன் ௴ ~1000 பேர் வந்தார்கள்

The above example clearly shows that the readability of the converted text is maintained and also at the same time loss of data during round-trip conversion is avoided.

Conclusion

Once the "Literal Start-End Toggle" character for each encoding and the "Standard Literal Equivalent" for the symbols are defined it will enable all developers of encoding conversion software to implement this procedure in an uniform manner and text can smoothly and flawlessly move between the various standard Tamil encoding schemes.

Author

P.Chellappan, Palaniappa Bros. email: chellappan@vsnl.com

TABLE-1

Symbol	Name	TAM	TAB	TSCII	UNICODE
‘	CP Open Single Quote*	212	212	--	--
’	CP Close Single Quote*	213	213	--	--
“	CP Open Double Quote*	210	210	--	--
”	CP Close Double Quote*	211	211	--	--
௳	Tamil Day Sign	115	--	--	U+0BF3**
௴	Tamil Month Sign	116	--	--	U+0BF4**
௵	Tamil Year Sign	117	--	--	U+0BF5**
௹	Tamil Ruppee Sign	150	--	--	U+0BF9**
௺	Tamil Number Sign	173	--	--	U+0BFA**
௶	Tamil Debit Sign	208	--	--	U+0BF6**
௷	Tamil Credit Sign	209	--	--	U+0BF7**
௸	Tamil As Above Sign	151	--	--	U+0BF8**
௧	Tamil Digit One	--	--	129	U+0BE7
௨	Tamil Digit Two	--	--	141	U+0BE8
௩	Tamil Digit Three	--	--	142	U+0BE9
௪	Tamil Digit Four	--	--	143	U+0BEA
௫	Tamil Digit Five	--	--	144	U+0BEB
ï	Tamil Digit Six	--	--	149	U+0BEC
ñ	Tamil Digit Seven	--	--	150	U+0BED
ó	Tamil Digit Eight	--	--	151	U+0BEE
ò	Tamil Digit Nine	--	--	152	U+0BEF
௰	Tamil Number Ten	--	--	157	U+0BF0
௱	Tamil Number Hundred	--	--	158	U+0BF1
ü	Tamil Number Thousand	--	--	159	U+0BF2
ெ0ா	Tamil Vowel Sign O	--	--	--	U+0BCA
ே0ா	Tamil Vowel Sign OO	--	--	--	U+0BCB
ெ0ள	Tamil Vowel Sign Au	--	--	--	U+0BCC

* - Cross Paltform (Mac to Win) compatibility characters

** - Proposed Draft Amendment (PDAM) 2 ISO/IEC 10646-1:2000/Amd.2:2001 (E)

TABLE - 2

Symbol	Literal Equivalents
‘	‘
’	’
“	“
”	”
௳	நாள்
௴	மாதம்
௵	வருடம்
௹	ரூபாய்
௺	எண்
௶	பற்று
௷	வரவு
௸	மேற்படி
௧	1
௨	2
௩	3
௪	4
௫	5
	6
ñ	7
ó	8
ò	9
௰	10
௱	100
ü	1000
ெ0ா	ெ0ா
ே0ா	ே0ா
ெ0ௗ	ெ0ௗ