Text conversion from TSCII 1.7 to Unicode

Written by:
Muthu Nedumaran
(Muthu@Murasu.Com)

Introduction

Unicode is the only industry standard for the Tamil script that is supported in commercial operating platforms. It is also fast becoming the preferred standard in open-source platforms like Linux.

However, before Unicode became prevalent, users of the Tamil script were building Websites, exchanging electronic mails and storing documents in an 8bit scheme called TSCII. The latest version of which is TSCII 1.7. Information on TSCII can be found at http://www.tamil.net/tscii.

TSCII incorporates all glyphs to render the Tamil script with the basic set of tables in a true-type font. No substitution or shaping tables were necessary. With just pair-wise kerning, a fairly decent Tamil font can produce text well enough for electronic and print publishing.

However, users have seen the value of moving to Unicode and some of the major sites are laying out plans for the migration. Almost all new sites that come up adopt Unicode to store and render their Tamil content.

This document is put together to help with the development of software tools to convert TSCII 1.7 based text to Unicode.

TSCII to Unicode

TSCII encodes glyphs and Unicode encodes characters. As such, there isn't a one-to-one mapping of code points for ALL characters. While vowels, consonants and numerals have a direct one-to-one mapping from TSCII to Unicode, compound characters must be converted from a string of TSCII glyphs. The table below lists the types of glyphs in TSCII and how they need to be handled in the conversion process:

TSCII Glyphs	Unicode Conversion Format
1. Independent vowels & Aytham	One-to-one mapping
2. Consonants	One-to-one mapping
3. Numerals	One-to-one mapping
4. Ligatures	Converts to a string of Unicode characters
5. Dependant vowels (Modifiers)	Post modifiers: one-to-one mapping Pre modifiers: one-to-one mapping but reordered Two part modifiers: no one-to-one mapping; must be handled as a string

Table 1: TSCII to Unicode conversion format for various groups of glyphs

Mapping Tables

The tables below provide detailed mapping of TSCII glyphs to Unicode characters

1. TSCII Independent Vowels and Aytham: One-to-one mapping

TSCII (Hex)	Unicode (Code point + Character name)	TSCII (Hex)	Unicode (Code point + Character name)
0xAB	U+0B85, Tamil Letter A	0xB2	U+0B8F, Tamil Letter EE
0xAC	U+0B86, Tamil Letter AA	0xB3	U+0B90, Tamil Letter AI
0xFE	U+0B87, Tamil Letter I	0xB4	U+0B92, Tamil Letter O
0xAE	U+0B88, Tamil Letter II	0xB5	U+0B93, Tamil Letter OO
0xAF	U+0B89, Tamil Letter U	0xB6	U+0B94, Tamil Letter AU
0xB0	U+0B8A, Tamil Letter UU		
0xB1	U+0B8E, Tamil Letter E	0xB7	U+0B83, Tamil Sign Visarga(Aytham)

Table 2: TSCII Independent Vowels to Unicode

2. TSCII Consonants: One-to-one mapping

TSCII (Hex)	Unicode (Code point + Character name)	TSCII (Hex)	Unicode (Code point + Character name)
0xB8	U+0B95, Tamil Letter KA	0xC3	U+0BB0, Tamil Letter RA
0xB9	U+0B99, Tamil Letter NGA	0xC4	U+0BB2, Tamil Letter LA
0xBA	U+0B9A, Tamil Letter CA	0xC5	U+0BB5, Tamil Letter VA
0xBB	U+0B9E, Tamil Letter NYA	0xC6	U+0BB4, Tamil Letter LLLA
0xBC	U+0B9F, Tamil Letter TTA	0xC7	U+0BB3, Tamil Letter LLA
0xBD	U+0BA3, Tamil Letter NNA	0xC8	U+0BB1, Tamil Letter RRA
0xBE	U+0BA4, Tamil Letter TA	0xC9	U+0BA9, Tamil Letter NNNA
0xBF	U+0BA8, Tamil Letter NA	0x83	U+0B9C, Tamil Letter JA
0xC0	U+0BAA, Tamil Letter PA	0x84	U+0BB7, Tamil Letter SSA
0xC1	U+0BAE, Tamil Letter MA	0x85	U+0BB8, Tamil Letter SA
0xC2	U+0BAF, Tamil Letter YA	0x86	U+0BB9, Tamil Letter HA

Table 3: TSCII Consonants to Unicode

3. TSCII Numerals: One-to-one mapping

TSCII (Hex)	Unicode (Code point + Character name)	TSCII (Hex)	Unicode (Code point + Character name)
0x80	U+0BE6, Tamil Digit Zero (*proposed for post 4.0 – see Note 1*)	0x95 0x96	U+0BEC, Tamil Digit Six U+0BED, Tamil Digit Seven
0x81	U+0BE7, Tamil Digit One	0x97	U+0BEE, Tamil Digit Eight
0x8D	U+0BE8, Tamil Digit Two	0x98	U+0BEF, Tamil Digit Nine
0x8E	U+0BE9, Tamil Digit Three	0x9D	U+0BF0, Tamil Number Ten
0x8F	U+0BEA, Tamil Digit Four	0x9E	U+0BF1, Tamil Number One Hundred
0x90	U+0BEB, Tamil Digit Five	0x9F	U+0BF2, Tamil Number One Thousand

Table 4: TSCII Numerals to Unicode

4. TSCII Ligatures: One glyph to a string of Unicode characters

TSCII ligatures can be divided into five groups:

- grantha ligatures
- mey series
- ukara series
- uukaara series
- 'di' and 'dii'

Except for grantha ligatures, the others can be converted by splitting the ligature into its base consonant and the associated dependant vowel.

Ligature$_{\text{TSCII}}$ = Consonant$_{\text{Unicode}}$ + Dependant_Vowel_Sign$_{\text{Unicode}}$

4.1 Grantha ligatures

There are three grantha ligatures, sri, ksha and ksh, which can be easily decomposed into a string of Unicode characters.

TSCII (Hex, Name)	Unicode (Code points)
0x82, SRI	U+0BB8 + U+0BCD + U+0BB0 + U+0BC0 *(*see note 2 below)*
0x87, KSHA	U+0B95 + U+0BCD + U+0BB7
0x8C, KSH	U+0B95 + U+0BCD + U+0BB7 + U+0BCD

Table 5: Grantha ligature substitution

4.2 Mey series

Mey_Ligature$_{\text{TSCII}}$ = Consonant$_{\text{Unicode}}$ + TAMIL_SIGN_VIRAMA (U+0BCD)

Note:

a. Virama, known as "Pulli" in Unicode 4.0, is not encoded in TSCII. There was no need for this character as all pulli-ligated glyphs were given separate code points.
b. The shaping of these ligatures in Unicode is handled in the font through substitution tables.

TSCII (Hex)	Unicode (Code points)	TSCII (Hex)	Unicode (Code points)
0xEC	U+0B95 + U+0BCD	0xF7	U+0BB0 + U+0BCD
0xED	U+0B99 + U+0BCD	0xF8	U+0BB2 + U+0BCD
0xEE	U+0B9A + U+0BCD	0xF9	U+0BB5 + U+0BCD
0xEF	U+0B9E + U+0BCD	0xFA	U+0BB4 + U+0BCD
0xF0	U+0B9F + U+0BCD	0xFB	U+0BB3 + U+0BCD
0xF1	U+0BA3 + U+0BCD	0xFC	U+0BB1 + U+0BCD
0xF2	U+0BA4 + U+0BCD	0xFD	U+0BA9 + U+0BCD
0xF3	U+0BA8 + U+0BCD	0x88	U+0B9C + U+0BCD
0xF4	U+0BAA + U+0BCD	0x89	U+0BB7 + U+0BCD
0xF5	U+0BAE + U+0BCD	0x8A	U+0BB8 + U+0BCD
0xF6	U+0BAF + U+0BCD	0x8B	U+0BB9 + U+0BCD

Table 6: Mey ligatures in TSCII decomposed to equivalent Unicode code points.

4.2 Ukara series

Ukara_Ligature$_{\text{TSCII}}$ = Consonant$_{\text{Unicode}}$ + TAMIL_VOWEL_SIGN_U (U+0BC1)

Note:

a. Only the non-grantha consonants are ukara-ligated in Tamil.
b. The shaping of these glyphs in Unicode is handled in the font through substitution tables.

TSCII (Hex)	Unicode (Code points)	TSCII (Hex)	Unicode (Code points)
0xCC	U+0B95 + U+0BC1	0xD5	U+0BB0 + U+0BC1
0x99	U+0B99 + U+0BC1	0xD6	U+0BB2 + U+0BC1
0xCD	U+0B9A + U+0BC1	0xD7	U+0BB5 + U+0BC1
0x9A	U+0B9E + U+0BC1	0xD8	U+0BB4 + U+0BC1
0xCE	U+0B9F + U+0BC1	0xD9	U+0BB3 + U+0BC1
0xCF	U+0BA3 + U+0BC1	0xDA	U+0BB1 + U+0BC1
0xD0	U+0BA4 + U+0BC1	0xDB	U+0BA9 + U+0BC1
0xD1	U+0BA8 + U+0BC1		Ukarams in grantha are rendered as the base grantha followed by the 'u' vowel sign in the Tamil script. As such, there are no ligatures for them.
0xD2	U+0BAA + U+0BC1		
0xD3	U+0BAE + U+0BC1		
0xD4	U+0BAF + U+0BC1		

Table 7: Ukara ligatures in TSCII decomposed to equivalent Unicode code points.

4.3 Uukaara series

Uukaara_Ligature$_{\text{TSCII}}$ = Consonant$_{\text{Unicode}}$ + TAMIL_VOWEL_SIGN_UU (U+0BC2)

Note:

a. Only the non-grantha consonants are uukaara-ligated in Tamil.
b. The shaping of these glyphs in Unicode is handled in the font through substitution tables.

TSCII (Hex)	Unicode (Code points)	TSCII (Hex)	Unicode (Code points)
0xDC	U+0B95 + U+0BC2	0xE5	U+0BB0 + U+0BC2
0x9B	U+0B99 + U+0BC2	0xE6	U+0BB2 + U+0BC2
0xDD	U+0B9A + U+0BC2	0xE7	U+0BB5 + U+0BC2
0x9C	U+0B9E + U+0BC2	0xE8	U+0BB4 + U+0BC2
0xDE	U+0B9F + U+0BC2	0xE9	U+0BB3 + U+0BC2
0xDF	U+0BA3 + U+0BC2	0xEA	U+0BB1 + U+0BC2
0xE0	U+0BA4 + U+0BC2	0xEB	U+0BA9 + U+0BC2
0xE1	U+0BA8 + U+0BC2		Uukaarams in grantha are rendered as the base grantha followed by the 'u' vowel sign in the Tamil script. As such, there are no ligatures for them.
0xE2	U+0BAA + U+0BC2		
0xE3	U+0BAE + U+0BC2		
0xE4	U+0BAF + U+0BC2		

Table 7: Uukaara ligatures in TSCII decomposed to equivalent Unicode code points.

4.4 'di' and 'dii'

These two ligatures can be simply substituted as follows:

TSCII	Unicode	TSCII	Unicode
0xCA	U+0B9F + U+0BBF	0xCB	U+0B9F + U+0BC0

Table 8: 'di' and 'dii' decomposed to equivalent Unicode code points

5. Dependant vowels (modifiers)

5.1 Post modifiers

Post modifiers in TSCII can be converted to Unicode using a straight one-to-one mapping.

TSCII (Hex)	Unicode (Code point)	Remarks
0xA1	U+0BBE	Straight mapping only works for aakaarams. If there is a 0xA6 or 0xA7 preceding the consonant before 0xA1, this is part of a two-part dependant vowel. (See Table 10 and Table 11)
0xA2	U+0BBF	
0xA3	U+0BC0	
0xA4	U+0BC1	
0xA5	U+0BC2	

Table 9: Post modifiers

5.2 Pre modifiers

As TSCII is a glyph encoding, pre-modifiers are placed before the base consonant. When converting to Unicode, these modifiers must be re-ordered: i.e. placed after the base consonant.

Pre_Modifier$_{\text{TSCII}}$ + Base_Consonant$_{\text{TSCII}}$ = Base_Consonant$_{\text{Unicode}}$ + Pre_Modifier$_{\text{Unicode}}$

TSCII (Hex)	Unicode (Code point)	Remarks
0xA6	U+0BC6	These modifiers are considered pre-modifiers as long as there is no 0xA1 following the consonant next to them (see table 9). Otherwise, they must be treated as two-part dependant vowels. (See Table 11)
0xA7	U+0BC7	
0xA8	U+0BC8	

Table 10: Pre-modifiers

5.3 Two-part vowels

In TSCII two modifiers are placed, one before and one after the base consonant. In Unicode, the text will contain the base consonant followed by the two-part dependant vowel. The shaping is taken care of by the font.

Pre_Modifier$_{\text{TSCII}}$ + Base_Consonant$_{\text{TSCII}}$ + Post_Modifier$_{\text{TSCII}}$ =
Base_Consonant$_{\text{Unicode}}$ + Two_Part_Dependant_Vowel$_{\text{Unicode}}$

In Table 11, *BCt* is used to denote Base_Consonant$_{\text{TSCII}}$ and *BCu* is used to denote Base_Consonant$_{\text{Unicode}}$. The mapping of *BCt* -> *BCu* is provided in Table 3.

TSCII (Hex)	Unicode (Code points)
0xA6 + *BCt* + 0xA1	*BCu* + U+0BCA
0xA7 + *BCt* + 0xA1	*BCu* + U+0BCB
0xA6 + *BCt* + 0xAA	*BCu* + U+0BCC

Table 11: Two-part vowel modifiers applied to base consonants

6. Other characters

TSCII 1.7 also includes five other characters that are not Tamil specific. They can be directly converted to their equivalent Unicode code points.

TSCII (Hex)	Unicode (Code point)	Unicode character name
0x91	U+2018	Left Single Quotation Mark
0x92	U+2019	Right Single Quotation Mark
0x93	U+201C	Left Double Quotation Mark
0x94	U+201D	Right Double Quotation Mark
0xA9	U+00A9	Copyright Sign

7. TSCII 1.6 considerations

The only difference between TSCII 1.6 and TSCII 1.7 is the hex value for Tamil Letter I. It was moved from 0xAD in 1.6 to 0xFE in 1.7. The conversion software may move 0xAD to 0xFE to "upgrade" the legacy text to TSCII 1.7 and then perform the conversion to Unicode. This way, text in TSCII 1.6 can also be converted to Unicode without errors.

Notes:

1. Code-point for TAMIL DIGIT ZERO has been proposed for inclusion beyond version 4.0. This is yet to be finalised.

2. A new code point for TAMIL LETTER SHA has been proposed for inclusion beyond version 4.0. As with TAMIL DIGIT 0, this character is yet to be finalised. Once approved, the new character may replace U+0BB8 in the formation of SRI. SHA is not present in TSCII.

References:

1. The TSCII 1.7 code chart: http://www.tamil.net/tscii/charset17.gif
2. Unicode 4.0, Tamil code chart: http://www.unicode.org/charts/PDF/U0B80.pdf

Date Created: 01 March 2004
Last update: 20 May 2004