TAM TO UNICODE CONVERSION

Author:  P.Chellappan
          Palaniappa Bros
          chellappan@vsnl.com

Introduction:

TAM is the official monolingual Tamil (8 bit) encoding scheme of the Government of Tamilnadu, which has the largest Tamil speaking population in the world. A vast amount of Tamil textual information in digital libraries, online newspapers, magazines etc., is available today in this encoding scheme. As Unicode is fast becoming the encoding by choice, there is a need for conversion from TAM encoded text to Unicode.

TAM, a monolingual Tamil encoding scheme, is a superset of the TAB bilingual Tamil encoding scheme.

TAM is a glyph encoding scheme, while Unicode is a character encoding scheme. Hence there exists a one-to-one, one-to-many, many-to-one or many-to-many relationship between the Tamil alphabets in TAM and those in Unicode.

A tamil alphabet in a TAM encoded text could be made up with a single, two or three bytes. The first part of this note describes, in a simple C like pseudo code, how to determine the string sequence in TAM that goes to make a Tamil alphabet. The second part provides a cross mapping table to convert this sequence into the corresponding Unicode string sequence.

Determine a Tamil alphabet:

vowel = { 0xDC, 0xDD, 0xDE, 0xDF, 0xE0, 0xE1, 0xE2, 0xE3, 0xE4, 0xE5, 0xE6 }
consonant = { 0xE8, 0xE9, 0xEA, 0xEB, 0xEC, 0xED, 0xEE, 0xEF, 0xF0, 0xF1, 0xF2, 0xF3, 0xF4, 0xF5, 0xF6, 0xF7, 0xF8, 0xF9, 0xFA, 0xFB, 0xFC, 0xFD, 0xFE }
grantha = { 0xFA, 0xFB, 0xFC, 0xFD, 0xFE }

while (not end of file)
{
  x = read(1)
  case (x = 0xE5)
  {
    y = read(1)
    if (y = 0xF7)
    {
      string = xy
    } else
    {
      string = x
      move(-1)
    }
  }
  convert (string to unicode)
  loop

  case (x = 0xAC)
  {
    y = read(1)
    if (y is a consonant)
    {
      string = xy
    } else
    {
      string = x
      move(-1)
    }
  }
  convert (string to unicode)
loop
}
case (x is a consonant)
{
    y = read(1)
    if (x is a grantha)
    {
        if (y = 0xA7 or y = 0xA8)  // 'U' or 'UU' modified grantha
            string = xy
              convert (string to unicode)
            loop
    }
    if (y = 0xA3)     // 'AA' modified consonant
                    string = xy
    else
    {
        string = x
        move(-1)
    }
    convert (string to unicode)
    loop
}
case (x = 0xAA)     // 'E' vowel modifier
{
    y = read(1)
    if (y is a consonant)
    {
        z = read(1)
        if (z = 0xA3 or z = 0xF7) // 'O' or 'AU' vowel modified consonant
            alphabet = xyz
        else // 'E' vowel modified consonant
            string = xy
        move(-1)
    }
    else // 'E' vowel modified consonant
    {
        string = x
        move(-1)
    }
    convert (string to unicode)
    loop
}
case (x = 0xAB)     // 'EE' vowel modifier
{
    y = read(1)
    if (y is a consonant)
    {
        z = read(1)
        if (z = 0xA3) // 'OO' vowel modified consonant
            alphabet = xyz
        else // 'EE' vowel modified consonant
            string = xy
        move(-1)
    }
    else // 'EE' vowel modified consonant
    {
        string = x
        move(-1)
    }
else {
    string = x
    move(-1)
}
convert (string to unicode string)
loop
}

Note :
read(1) means read one byte
move(-1) means move the file pointer back one byte
## The entries in the second section are in the Tamil alphabetical sequence
and contain the many-to-one and many-to-many relationships.
The code points that make up the string are separated by a space.

A range of code points are indicated by a hyphen between the start and the end code points

If a TAM code point is not defined or it has no equivalent UNICODE code point
the UNICODE code point is left blank.

Control character mappings are not shown in this table, following
the conventions of the standard UTC mapping tables. However, the
TAM character set uses the standard control characters at
0x00-0x1F and 0x7F.

### SECTION ONE - (one-to-one and one-to-many relationship)

<table>
<thead>
<tr>
<th>Code Point</th>
<th>Tamil Character</th>
</tr>
</thead>
<tbody>
<tr>
<td>0x0020</td>
<td>SPACE</td>
</tr>
<tr>
<td>0x0021</td>
<td>EXCLAMATION MARK</td>
</tr>
<tr>
<td>0x0022</td>
<td>QUOTATION MARK</td>
</tr>
<tr>
<td>0x0023</td>
<td>NUMBER SIGN</td>
</tr>
<tr>
<td>0x0024</td>
<td>DOLLAR SIGN</td>
</tr>
<tr>
<td>0x0025</td>
<td>PERCENT SIGN</td>
</tr>
<tr>
<td>0x0026</td>
<td>AMPERSAND</td>
</tr>
<tr>
<td>0x0027</td>
<td>APOSTROPHE</td>
</tr>
<tr>
<td>0x0028</td>
<td>LEFT PARENTHESES</td>
</tr>
<tr>
<td>0x0029</td>
<td>RIGHT PARENTHESES</td>
</tr>
<tr>
<td>0x002A</td>
<td>ASTERISK</td>
</tr>
<tr>
<td>0x002B</td>
<td>PLUS SIGN</td>
</tr>
<tr>
<td>0x002C</td>
<td>COMMA</td>
</tr>
<tr>
<td>0x002D</td>
<td>HYPHEN-MINUS</td>
</tr>
<tr>
<td>0x002E</td>
<td>FULL STOP</td>
</tr>
<tr>
<td>0x002F</td>
<td>SOLIDUS</td>
</tr>
<tr>
<td>0x0030</td>
<td>DIGIT ZERO</td>
</tr>
<tr>
<td>0x0031</td>
<td>DIGIT ONE</td>
</tr>
<tr>
<td>0x0032</td>
<td>DIGIT TWO</td>
</tr>
<tr>
<td>0x0033</td>
<td>DIGIT THREE</td>
</tr>
<tr>
<td>0x0034</td>
<td>DIGIT FOUR</td>
</tr>
<tr>
<td>0x0035</td>
<td>DIGIT FIVE</td>
</tr>
<tr>
<td>0x0036</td>
<td>DIGIT SIX</td>
</tr>
<tr>
<td>0x0037</td>
<td>DIGIT SEVEN</td>
</tr>
<tr>
<td>0x0038</td>
<td>DIGIT EIGHT</td>
</tr>
<tr>
<td>0x0039</td>
<td>DIGIT NINE</td>
</tr>
<tr>
<td>0x003A</td>
<td>COLON</td>
</tr>
<tr>
<td>0x003B</td>
<td>SEMICOLON</td>
</tr>
<tr>
<td>0x003C</td>
<td>LESS-THAN SIGN</td>
</tr>
<tr>
<td>0x003D</td>
<td>EQUALS SIGN</td>
</tr>
<tr>
<td>0x003E</td>
<td>GREATER-THAN SIGN</td>
</tr>
<tr>
<td>0x003F</td>
<td>QUESTION MARK</td>
</tr>
<tr>
<td>0x0040</td>
<td>COMMERCIAL AT</td>
</tr>
<tr>
<td>0x0B95</td>
<td>TAMIL LETTER KI</td>
</tr>
<tr>
<td>0x0B99</td>
<td>TAMIL LETTER NGI</td>
</tr>
<tr>
<td>0x0B9A</td>
<td>TAMIL LETTER CI</td>
</tr>
<tr>
<td>0x0B9E</td>
<td>TAMIL LETTER NYI</td>
</tr>
<tr>
<td>0x0BA3</td>
<td>TAMIL LETTER NNI</td>
</tr>
<tr>
<td>0x0BA4</td>
<td>TAMIL LETTER TI</td>
</tr>
<tr>
<td>0x0BA8</td>
<td>TAMIL LETTER NI</td>
</tr>
<tr>
<td>0x0BAA</td>
<td>TAMIL LETTER PI</td>
</tr>
<tr>
<td>0x0BAE</td>
<td>TAMIL LETTER MI</td>
</tr>
<tr>
<td>0x0BAF</td>
<td>TAMIL LETTER YI</td>
</tr>
<tr>
<td>0xBB0</td>
<td>TAMIL LETTER RI</td>
</tr>
<tr>
<td>0xBB2</td>
<td>TAMIL LETTER LI</td>
</tr>
<tr>
<td>0xBB5</td>
<td>TAMIL LETTER VI</td>
</tr>
</tbody>
</table>
# SECTION TWO - (many-to-one and many-to-many relationship)
# -------------------------------------------------------------
# 0xE5 0xF7  0x0B94  # TAMIL LETTER AU
# 0xE8 0xA3  0x0B95 0x0BBE  # TAMIL LETTER KAA
# 0xE9 0xA3  0x0BB9 0x0BBE  # TAMIL LETTER NGAA
# 0xEA 0xA3  0x0BB9 0x0BBE  # TAMIL LETTER CAA
# 0xEB 0xA3  0x0BB9 0x0BBE  # TAMIL LETTER NYAA
# 0xEC 0xA3  0x0BB9 0x0BBE  # TAMIL LETTER TAA
# 0xF6 0xA3  0x0BB9 0x0BBE  # TAMIL LETTER NAA
# 0xF7 0xA3  0x0BB9 0x0BBE  # TAMIL LETTER YAA