Malayalam cillaksarams

Eric Muller, Adobe Systems Inc.
May 14, 2006

1. Introduction
2. Sources
3. Signs for pure consonants (cillaksaram)
4. Chandrakkala
5. ISCII
6. TUS 4.0
7. Problematic situations
8. Encoding the cillaksarams
9. Alternatives
Document History

1. Introduction

This document is a revision of L2/05-148, which incorporates additional evidence, and proposes to encode the cillaksaram.

The following individuals have participated in the original discussion on the indic@unicode.org list: Omi Azad, Vinod Balakrishnan, Stefan Baums, Peri Bhaskararao, Varghese Chacko, Gihan Dias, Ketaki Kushari Dyson, Micheal Everson, Soleiman Karim, Nishad Kaypally, Jonathan Kew, Antoine Leca, Rick McGowan, Mike Meir, Eric Muller, Paul Nelson, Mahesh T. Pai, Hariram Pansari, Dr. U. B. Pavanaja, Rajkumar S., Deepayan Sarkar, Rajeev J. Sebastian, Gautam Sengupta, Sukhjinder Sidhu, Steve Smith, Kevin Sooryan, Sinnathurai Srivas, K. G. Sulochana, Owen Taylor, Anirban Udr, Uma Umamaheswaran, Ken Whistler.

This document also benefited from the contributions submitted to the UTC listed in the next section.

2. Sources

[TDIL] Malayalam, TDIL newsletter, April 2002, available at http://tdil.mit.gov.in/news.htm.

[Keralapanineeyam] Kerala Panineeyam, A. R. Rajaraja Varma, available at http://www.malayalamresourcecentre.org/Mrc/literature/keralapaanineeyam/panineeyam.html.

[Frohnmeyer] A progressive grammar of the Malayalam language, L. Johannes Frohnmeyer, second edition, Basel Mission, 1913. Reprinted by Asian Educational Services, 1979. Scans of some of the pages are available at http://www.unicode.org/~emuller/iwg/sources/pl4713.f7_1979/index.html.

[Mohanan] Malayalam Writing, K. P. Mohanan, in The World's Writing Systems, edited by Peter T. Daniels and William Bright, Oxford University Press, 1996.

[L2/05-081] Chilling Effects of the Chillu, Mahesh Pai. Also available at http://paivakil.port5.com/writings/chill_effect.shtml.

[L2/05-085], Encoding of Chillu forms in Malayalam (PRI #66 feedback), Cibu C. Johny

[L2/05-148] Comments on PRI 66: Malayalam cillaksarams, Eric Muller

[L2/05-154], Malayalam chillu ka, Peter Constable.

[L2/05-210] Chandrakkala. Samvruthokaram. Chillaksharam, from the perspective of Malayalam Collation, R. Chitrajakumar , N. Gangadharan

[L2/05-213] Samvruthokaram and Chandrakkala, R. Chitrajakumar , N. Gangadharan

[L2/05-214] Chillaksharam of Malayalam Language, R. Chitrajakumar , N. Gangadharan

[L2/05-236] Malayalam language -- inclusion of chillu characters, Kerala State IT Mission

[L2/05-246] Letter to Mark Davis from Om Vikas re Malayalam Chillus, Om Vikas

[L2/05-307] ZWJ/ZWNJ behavior under Indic scripts with special reference to chillu, conjuncts, etc in Malayalam, Rajeev J Sebastian

[L2/05-308] Problems of Malayalam Encoding in the Indic context (Rachana's response to Malayalam encoding debate), R. Chitrajakumar , N. Gangadharan

[L2/05-310] Rachana Documents (cover letter for L2/05-307, 308, 309), R. Chitrajakumar

[L2/05-334] Critical review of Rachana (L2/05-210) and other arguments to encode Malayalam Chillus, Cibu C Johny

[L2/05-354] Chillu and Semivowel examples (this is an addendum to L2/05-308), R. Chitrajakumar , N. Gangadharan

[L2 05/372] Chillus, Samvrithokaram and Chandrakkala - A Problem Which is Not, Mahesh T. Pai

[L2/05-373] Malayalam keyboard layout and character encoding. Report of the Government of Kerala Committee, May 2001, Government of Kerala. Also available at http://www.keralaitmission.org/malayalam/malayalam_keybo.htm and at http://www.malayalamresourcecentre.org/Mrc/report.pdf

[L2/06-189] The chillu issue, Government of India, MIT

[L2/06-190] Annexure 1 - Chillu evidence, Government of India, MIT

3. Signs for pure consonants (cillaksaram)

§1. Malayalam uses some signs to write pure consonants. Those signs are called *cillaksaram*. For example, ൽ is the sign for the pure consonant corresponding to ല. For this document, ല is called the underlying consonant of ൽ.

§2. Various sources show different sets, as well as different/multiple underlying consonants. Here is a cumulative summary:

full consonant	na	nna	la	ra	rra	lla	zha	ka
	ന	ണ	ല	ര	റ	ള	ഴ	ക
cillaksaram	ൻ	ൺ	ൽ	ർ		ൾ		ൿ
Kerala Panineeyam	x	x	x	x	x	x	x	
Frohnmeyer	x	x	x	x		x		
Mohanan	x	x	x	x		x		
ISCII 91	x	x	x	x		x		
Unicode 4.0	x	x	x		x	x		

The Panineeyam indicates that some cillaksarams have two possible underlying consonants.

While the anusvara ം and the visarga ഃ represent pure consonants (corresponding to മ and ഹ), and the Kerala Panineeyam considers the anusvara ം as just another cillaksarams, this is not the prevalent point of view.

Both Frohnmeyer and Mohanan use an alternate form for the lla cillaksaram, which is also given in Keralapanineeyam. This seems to be an historic glyphic variant.

§3. There is reportedly a cillaksaram for ൐, but we have not been able to find a reliable source for it, nor any image.

§4. In modern Malayalam, only the first five cillaksaram are considered common. However, there is enough evidence of the existence of the ൿ cillaksaram (e.g. in L2/05-154) to include it as well, should the cillaksarams be encoded.

§5. It is worth noting that for sorting purposes, the Kerala IT Mission document sorts the cillaksaram just before their underlying consonants, with ൻ just before ന and after ര, ൾ just before ഴ, the anusvara just before മ and the visarga ഃ just before ഹ.

4. Chandrakkala

§6. Malayam has a “half-u” sound, often at the end of words. The exact pronunciation of this sound varies according to region, but it is approximately [ə].

§7. According to Frohnmeyer, there are two traditions to write this sound. The northern tradition is to use the sign ◌ു, known

as the chandrakkala, and this sign functions just like a vowel sign. The southern tradition is to use both the vowel sign U and the chandrakkala. This vowel sign U is called *samvrittokarama*. Since Frohnmeyer (early 20th century), this southern tradition is less commonly used, even in the south.

§8. Mohanan in fact lists the half-u vowel along with the other vowels (see Table 38.1, p421), indicates that there is no independent symbol for it, and gives the chandrakkala as the diacritic sign.

§9. Malayalam also uses the chandrakkala to mark a consonant stripped of its inherent vowel (i.e. as a visible virama), whenever the representation of a consonant cluster does not have a ligated or conjunct form. (As usual in the Indic scripts, the set of conjunct forms is font/style dependent).

The script reform, which tends to deprecate the use of a large number of consonant conjunct signs, encourages the increased use of the chandrakkala to write pure consonant sounds in conjuncts. However, there are pre-reform occurrences of the chandrakkala used to write pure consonants, because some consonant clusters did not have a conjunct form.

5. ISCII

§10. ISCII 91 spells out that cillaksarams are represented using the soft halant, i.e. by a combination of <base consonant, HALANT, NUKTA>. Note that the use of the nukta character is not related to the nukta sign: the convention is for the sequence as a whole.

6. TUS 4.0

§11. TUS 4.0 indicates that the sequence <consonant, VIRAMA, ZWJ> is used to represent the cillaksaram, where "consonant" is the underlying consonant for the cillaksaram. In general, this pattern is used to represent the half-form fo the consonant.

Furthermore, TUS 4.0 lists five cillaksram, and takes the position that RRA is the underlying consonant of ൻ and LLA is the underlying consonant of ൽ.

§12. TUS 4.0 indicates that the sequence <consonant, VIRAMA, ZWNJ> is used to represent a pure consonant using the chandrakkala.

§13. TUS 4.0 does not provide any means of distinguishing in the representation of text whether a consonant with a chandrakkalla represent the pure form of that consonant or that consonant and a half-u vowel.

7. Problematic situations

§14. TUS 4.0 does not clearly indicate how the sequence <consonant, VIRAMA> should be rendered, when the consonant does not form a conjunct with a following consonant and a cillaksaram is possible. Some readers have concluded that this sequence is always rendered with a chandrakkala, others that it is rendered with a cillaksaram if one exists, with a chandrakkala otherwise.

§15. In L2/05-334, Cibu provides a number of examples of the form:

വൻയവനിക *van_yavanika*, big curtain

വന്യവനിക *vanyavanika*, wild forest

In the first line, we have a ൻ cillaksaram followed by a യ *ya*; in the second line we have the consonant ന *na* underlying that cillaksaram which enters in a conjunct with the following *ya* (in this particular conjunct, the *ya* takes an alternate form).

Under the current Unicode model, those two words are represented by sequences that differ only in a joiner: <0D35 VA, 0D328 NA, 0D4D VIRAMA, 200D ZWJ, 0D2F YA, ...> for the first, and the same sequence without a zwj for the second. Since those words are distinct and their written forms are not interchangeable, this does not follow the general pattern of Indic script encoding, in which the joiners affect only the preferred rendering, without comprimising the legibility of the text.

While one could think of changing the ignorable status of joiners, this is now becoming more difficult, as the IDN machinery has already adopted the ignorable status. In other words, it's not just Unicode we need to change, but IDN as well.

§16. The report of the Kerala IT mission and the TDIL document derived from it show this contrast:

എന്റെ */ente/*

ഹെൻറി */henry/*

where the ◌് sign is either below or next to the ൻ cillaksaram. When below, the ◌് represents the sound /t/. When after, it represents its usual sound /r/.

Under the current model, it is natural to represent the second word by <..., 0D28 NA, 0D4D VIRAMA, 200D ZWJ, 0D31 RRA, ...>. However, the representation of the first word is problematic.

It is arguable that the difference, from a written point of view, is that the ൻ and the ന are part of the same orthographic cluster in the first case, and part of two successive clusters in the second case. This is normally represented by the introduction of a virama between them when they are in the same cluster. However, the representation of the ൻ cillaksaram already uses a virama character, so the result would be a sequence with two viramas in a row, a pattern that would be new in Unicode.

To avoid this, the only alternative (without encoding the cillaksarams) is to use some sort of joiner to distinguish the sequences, but we then run into the previous problem, since joiners would then no longer be ignorable.

§17. L2/05-081 shows examples of contrast between a half-u sound and a pure consonant sound. In those examples, which are fairly common in practice, the ability to distinguish between the two situations depends critically on the use of cillaksaram to write the pure consonant sound:

ആ മനുഷ്യൻ കൊടുക്കുന്നു that man is giving.

ആ മനുഷ്യന് കൊടുക്കുന്നു giving to the man.

Again, we have a situation of two non-interchangeable orthographic forms which differ only in the use of a joiner.

8. Encoding the cillaksarams

§18. Without denying the relationship between a cillaksaram and its underlying consonant, the examples above show that from the graphic point of view, they are used constrastively, and that the current representation makes the joiners semantically significant. Because of the general principle on joiners, and the application of this principles in environments such as IDNs, the current representation is problematic. In addition, there is uncertainty on the consonant underlying some cillaksarams.

§19. To overcome these problems, the proposal is to encode the cillaksarams:

Proposed code point	Representative glyph	Name
0D7A	ൺ	MALAYALAM LETTER NN
0D7B	ൻ	MALAYALAM LETTER N
0D7C	ർ	MALAYALAM LETTER RR
0D7D	ൽ	MALAYALAM LETTER L
0D7E	ൾ	MALAYALAM LETTER LL
0D7F	ൿ	MALAYALAM LETTER K

The remaining properties are the same as the other consonants: jamo= gc=Lo ccc=0 dt=no dm= nt=no nv= bc=L Bidi_M=n bmg= suc= slc= stc= uc= lc= tc= ccf= scf= fcf= tcf= jt=U jg= ea=N lb=AL sc=Mlym Dash=n WSpace=n Hyphen=n QMark=n Radical=n Ideo=n UIdeo=n IDSB=n IDST=n hst=na DI=n ODI=n Alpha=y OAlpha=n Upper=n OUpper=n Lower=n OLower=n Math=n OMath=n Hex=n AHex=n NChar=n VS=n Bidi_C=n Join_C=n Gr_Base=y Gr_Ext=n OGr_Ext=n Gr_Link=n STerm=n Ext=n Term=n Dia=n Dep=n IDS=y OIDS=n XIDS=y IDC=y OIDC=n XIDC=y SD=n LOE=n Pat_WS=n Pat_Syn=n GCB=Other WB=ALetter SB=OLetter CE=n Comp_Ex=n NFC_QC=y NFD_QC=y NFKC_QC=y NFKD_QC=y XO_NFC=n XO_NFD=n XO_NFKC=n XO_NFKD=n FC_NFKC= isc= na1=.

§20. This solves the four problems above:

- §14: <consonant, VIRAMA>, with a non-cillaksaram consonant, is always rendered with chandrakkala

- §15: the first word is <0D35 VA, 0D7B N, 0D2F YA, ...>, the second is <0D35 VA, 0D28 NA, 0D4D VIRAMA, 0D2F YA, ...>.
- §16: the first word is <0D0E E, 0D7B N, 0D4D VIRAMA, 0D31 RRA, 0D46 SIGN E>; the second is <0D39 HA, 0D46 SIGN E, 0D7B N, 0D31 RRA, 0D3F I>.
- §17: the first word is <...0D7B N, 0020 SPACE, ...>, the second is <...0D28 NA, 0D4D VIRAMA, 0020 SPACE>.

§21. For compatibility with previous versions of Unicode, the best which can be done is to interpret sequences of the form <consonant, 0D4D VIRAMA, 200D ZWJ>, where consonant is one of U+0D23 MALAYALAM LETTER NNA, U+0D28 MALAYALAM LETTER NA, U+0D31 MALAYALAM LETTER RRA, U+0D32 MALAYALAM LETTER LA, U+0D33 MALAYALAM LETTER LLA as equivalent to the corresponding new character. This is viable because U+200D ZERO WIDTH JOINER is currently rarely used (if at all) for another purpose than to represent a cillaksaram.

9. Alternatives

§22. Without encoding the cillaksarams, there is apparently no alternative that does not involve:

- make the joiners semantically significant, and thereby exclude a number of words from IDNs as they are defined today, and
- instituting new patterns of character sequences, either involving two VIRAMA characers in a row or a refinenment of the use of joiners.

Document History

Author: Eric Muller

Revision	Date	Comments
2	May 14, 2006	Major revision to include the documents submitted to the UTC.
Revision	Date	Comments
1	May 11, 2005	Initial version

ISO/IEC JTC 1/SC 2/WG 2
PROPOSAL SUMMARY FORM TO ACCOMPANY SUBMISSIONS
FOR ADDITIONS TO THE REPERTOIRE OF ISO/IEC 10646[1]
Please fill all the sections A, B and C below.
Please read Principles and Procedures Document (P & P) from http://www.dkuug.dk/JTC1/SC2/WG2/docs/principles.html **for guidelines and details before filling this form.**
Please ensure you are using the latest Form from http://www.dkuug.dk/JTC1/SC2/WG2/docs/summaryform.html**.**
See also http://www.dkuug.dk/JTC1/SC2/WG2/docs/roadmaps.html **for latest *Roadmaps*.**

A. Administrative

1. **Title:** *Malayalam cillaksarams*
2. Requester's name: *Unicode Consortium*
3. Requester type (Member body/Liaison/Individual contribution): *Liaison*
4. Submission date: *May 14, 2006*
5. Requester's reference (if applicable):
6. Choose one of the following:
   This is a complete proposal: *Yes*
   (or) More information will be provided later:

B. Technical – General

1. Choose one of the following:
   a. This proposal is for a new script (set of characters):
      Proposed name of script:
   b. The proposal is for addition of character(s) to an existing block: *Yes*
      Name of the existing block: *Malayalam*
2. Number of characters in proposal: *6*
3. Proposed category (select one from below - see section 2.2 of P&P document):
   A-Contemporary X B.1-Specialized (small collection) B.2-Specialized (large collection)
   C-Major extinct D-Attested extinct E-Minor extinct
   F-Archaic Hieroglyphic or Ideographic G-Obscure or questionable usage symbols
4. Proposed Level of Implementation (1, 2 or 3) (see Annex K in P&P document): *1*
   Is a rationale provided for the choice?
      If Yes, reference:
5. Is a repertoire including character names provided? *Yes*
   a. If YES, are the names in accordance with the "character naming guidelines" in Annex L of P&P document? *Yes*
   b. Are the character shapes attached in a legible form suitable for review? *Yes*
6. Who will provide the appropriate computerized font (ordered preference: True Type, or PostScript format) for publishing the standard? *Unicode Consortium*
   If available now, identify source(s) for the font (include address, e-mail, ftp-site, etc.) and indicate the tools used:
7. References:
   a. Are references (to other character sets, dictionaries, descriptive texts etc.) provided? *Yes*
   b. Are published examples of use (such as samples from newspapers, magazines, or other sources) of proposed characters attached? *No*
8. Special encoding issues:
   Does the proposal address other aspects of character data processing (if applicable) such as input, presentation, sorting, searching, indexing, transliteration etc. (if yes please enclose information)? *No*

9. Additional Information:

Submitters are invited to provide any additional information about Properties of the proposed Character(s) or Script that will assist in correct understanding of and correct linguistic processing of the proposed character(s) or script. Examples of such properties are: Casing information, Numeric information, Currency information, Display behaviour information such as line breaks, widths etc., Combining behaviour, Spacing behaviour, Directional behaviour, Default Collation behaviour, relevance in Mark Up contexts, Compatibility equivalence and other Unicode normalization related information. See the Unicode standard at http://www.unicode.org for such information on other scripts. Also see http://www.unicode.org/Public/UNIDATA/UCD.html and associated Unicode Technical Reports for information needed for consideration by the Unicode Technical Committee for inclusion in the Unicode Standard.

[1] Form number: N3002-F (Original 1994-10-14; Revised 1995-01, 1995-04, 1996-04, 1996-08, 1999-03, 2001-05, 2001-09, 2003-11, 2005-01, 2005-09, 2005-10)

C. Technical - Justification

1. Has this proposal for addition of character(s) been submitted before? *No*
   If YES explain
2. Has contact been made to members of the user community (for example: National Body, user groups of the script or characters, other experts, etc.)? *Yes*
   If YES, with whom? *Government of India*
   If YES, available relevant documents: *References included*
3. Information on the user community for the proposed characters (for example: size, demographics, information technology use, or publishing use) is included? *No*
   Reference:
4. The context of use for the proposed characters (type of use; common or rare) *Common*
   Reference:
5. Are the proposed characters in current use by the user community? *Yes*
   If YES, where? Reference: *Kerala State, India*
6. After giving due considerations to the principles in the P&P document must the proposed characters be entirely in the BMP? *Yes*
   If YES, is a rationale provided? *No*
   If YES, reference:
7. Should the proposed characters be kept together in a contiguous range (rather than being scattered)? *Yes*
8. Can any of the proposed characters be considered a presentation form of an existing character or character sequence? *No*
   If YES, is a rationale for its inclusion provided?
   If YES, reference:
9. Can any of the proposed characters be encoded using a composed character sequence of either existing characters or other proposed characters? *No*
   If YES, is a rationale for its inclusion provided?
   If YES, reference:
10. Can any of the proposed character(s) be considered to be similar (in appearance or function) to an existing character? *Yes*
    If YES, is a rationale for its inclusion provided? *Yes*
    If YES, reference: *See attached*
11. Does the proposal include use of combining characters and/or use of composite sequences? *No*
    If YES, is a rationale for such use provided?
    If YES, reference:
    Is a list of composite sequences and their corresponding glyph images (graphic symbols) provided?
    If YES, reference:
12. Does the proposal contain characters with any special properties such as control function or similar semantics? *No*
    If YES, describe in detail (include attachment if necessary)
13. Does the proposal contain any Ideographic compatibility character(s)? *No*
    If YES, is the equivalent corresponding unified ideographic character(s) identified?
    If YES, reference: