Report on the Issues of Malayalam Language in Unicode

Version	Author	Date
0.1	Santhosh Thottingal	1-November-2011

Introduction

This is report on the existing character encoding issues in Malayalam Language(ml_IN). This report explains the history of key decisions by UTC and the impact to the Malayalam language content space and issues created by those decisions. The report does not explain the linguistic , technical arguments raised against such decisions, but the timeline and documents of those discussions. The intention of this report is to bring UTC's attention to the issues and requesting solutions.

History of the issues

Brief History of The Chillu issue[1]:

In 2004 Government of India's IT ministry recommended to atomically encode the Chillu characters of Malayalam. UTC issued a Public Review item (PR 66) requesting input on the question of encoding atomic chillu characters. Also, the topic was raised on the Unicode Indic list. Based on the responses to PR 66, UTC decided to encode the Chillu characters in UTC meeting in May 2005.

The following documents were presented in response to PRI 66[2].

1. L2/05-081 "Chilling Effects of the Chillu", Mahesh T. Pai, March 14, 2005. This is a snapshot on that date of http://paivakil.port5.com/writings/chill_effect.shtml.
2. L2/05-085 "Encoding of Chillu Forms in Malayalam", Cibu C Johny, May 2, 2005.
3. L2/05-148 "Comments on PRI 66: Malayalam cillaksaram", Eric Muller, May 11, 2005. This document is a summary of what was discussed on the indic@unicode.org list up to that point.
4. In addition, comments received via the web form were collected (along with comments related to over PRIs) in L2/05-111. Here is the part concerning PRI 66.

Just before the next UTC meeting, Rachana Aksharavedi (a group of people who came together to preserve Malayalam traditional script, which was being replaced by reformed script after the script reforms in 1981) submitted documents[3] presenting arguments against encoding. This indicated to UTC that the responses to PR 66 were not representative of the entire user community and may not have fully considered all the

[1] As narrated by Peter Constable at Indic@unicode.org mailing list - History of UTC activity related to encoding atomic chillu characters Wed 13/12/2006

[2] Documents related to PR66 -archived by Eric Muller- http://unicode.org/~emuller/iwg/p28/

[3] "Chandrakkala. Samvruthokaram. Chillaksharam." by R. Chitrajakumar and N. Gangadharan http://unicode.org/~emuller/iwg/p28/05210-malayalam.pdf

arguments, and so UTC took the extraordinary action of rescinding their previous decision so as to allow further investigation and evaluation.

More responses on this item came to Unicode’s Indic mailing list, and heated debates ensued.

In 2006, Eric Muller prepared a document summarizing the issues and the significant arguments in favour of encoding -- these constituted issues that (assuming the data was factual) the current encoded representation did not seem to handle well. On that basis, and with the Government of India's on-going request, UTC agreed once again to encode these characters (May 2006).

At the same time, UTC was also aware that some Malayalees remained opposed to encoding these characters, and that some of those had indicated they were going to work with other stakeholders in India, including the government, towards arriving at a consensus. Hence, UTC once more took action to delay the encoding process by having the characters moved, within the ISO process, from amendment 3 of ISO 10646 to amendment 4.
(Reference: as narrated by Peter Constable at Indic@unicode.org Mailing list (History of UTC activity related to encoding atomic Chillu characters Wed 13/12/2006 9:32 PM)

After a long, heated discussion in Unicode mailing list(can be read from 2007 archives of indic mailing list), UTC decided to go ahead with encoding Atomic Chillus. In unicode 5.1, released on April 2008, The new atomic code points we introduced.

Table 1 - The Chillus In Malayalam.

	Visual	**Representation in 5.0 and Prior**	**Preferred 5.1 Representation**
1	ൺ	NNA, VIRAMA, ZWJ (0D23, 0D4D, 200D)	0D7A MALAYALAM LETTER CHILLU NN
2	ൻ	NA, VIRAMA, ZWJ (0D28, 0D4D, 200D)	0D7B MALAYALAM LETTER CHILLU N
3	ർ	RA, VIRAMA, ZWJ (0D30, 0D4D, 200D)	0D7C MALAYALAM LETTER CHILLU RR
4	ൽ	LA, VIRAMA, ZWJ (0D32, 0D4D, 200D)	0D7D MALAYALAM LETTER CHILLU L
5	ൾ	LLA, VIRAMA, ZWJ (0D33, 0D4D, 200D)	0D7E MALAYALAM LETTER CHILLU LL
6	ൿ	*undefined*	0D7F MALAYALAM LETTER CHILLU K

Brief History of The NTA issue:

In Unicode version 5.1.0, UTC recommended[1] to use CHILLU N + VIRAMA+ RRA for ന്റ. This set a wrong precedence wherein Unicode dictated how a character sequence was written instead of just defining the code points. Since the existing NTA formed by NA + VIRAMA+ RRA did not cause any issues, every body continued to use it. Meanwhile Microsoft's Kartika font had a bug in rendering this sequence[2] (along with many other rendering issues). This was the default font for Malayalam locale in Microsoft's Windows XP and created a lot of confusion among users, input tool writers and this issue was finally brought to UTC. Using a VIRAMA after CHILLU is against all language rules of Malayalam. No rendering engine supports this kind of rendering and no Malayalam typographers want to make this mistake. Just canceling this remark in standard would be fine since nobody cared about that remark so far.

Brief History of The AU Vowel sign issue:

ൌ-sign double encoding problem: Unicode characters <U+0D4C> and <U+0D57> are interchangeably used in current day digital Malayalam though there is no canonical equivalence defined by the Unicode standard . U+04DC is equivalent to U+0D46 MALAYALAM VOWEL SIGN E + ◌ൗ U+0D57 MALAYALAM AU LENGTH MARK. Interchanging one part or two part signs does not affect the meaning of any Malayalam word. This fact proves that this is a formatting difference - not a spelling difference. ൌ-sign is spelled as a single entity in traditional as well as modern Malayalam. So it should have had only one encoding in Unicode. Since it already has two encoding, an equivalence definition is a must to avoid the dual encoding.

The public review issue #93 examines various options to avoid this dual encoding[3]. But both codepoints exist in Unicode and dual encoding issue is still present as of this writing.

Impact and Current state

Chillu

Both atomic and ZWJ based Chillu characters are equally used in present day digital Malayalam. Major news paper portals of Malayalam - Mathrubhumi, Mangalam, Deshabhimani, Madhyamam uses ZWJ based Chillu characters. Malayalam interface of Google applications uses ZWJ based Chillus. The widely used transliteration based input tool from Google produce ZWJ based Chillu[4]. In social media sites, both kinds of Chillu letters are used.

[1] New characters in Unicode 5.1.0 version - http://unicode.org/versions/Unicode5.1.0/#Malayalam_Chillu_Characters

[2] Discussion in Indic mailing list of Unicode, 2009 http://www.unicode.org/~ecartis/indic/indic.2009-10

[3] Public Review Issue #93: Representation of Malayalam /au/ Vowel in Traditional and Modern Orthography http://www.unicode.org/review/pr-93.html

[4] Google transliteration tool: http://www.google.com/transliterate

The Inscript typing standard, which is taught in Schools of Kerala uses ZWJ based Chillu[1]. Same Inscript keyboard is used for Malayalam Computing training all over Kerala.

Malayalam wikipedia[2], with nearly 20000 articles, uses atomic Chillu. The user input is passed through a normalization system to translate all ZWJ based Chillus to corresponding atomic Chillus before reaching mediawiki database. But this force conversion also causes many issues since it causes external links to sites with ZWJ based chillus break[3].

The GNU/Linux Operating System Distributions does not ship atomic Chillu compatible fonts except less popular Lohit Malayalam. The default input tools(inscript, transliteration based. phonetic) does not support inputting Atomic Chillu. This is largely because of the Free Software Community developers' disagreement with the the dual encoding and measure to avoid breaking backward compatibility.

One of the popular Malayalam font - Anjali Old Lipi has both kind of Chillus and thereby users are hidden from the data difference. Microsoft’s Kartika, the default font of Windows operating systems follow the same path.

Windows 7 does not have the facility to type atomic Chillu while the default keyboard support ZWJ based Chillu character[4].

Since there is a lot of digital content present in both encoding now, the search results from search engines like Google is unreliable unless one search for a word in both version of encoding.

Since the search results with both kind of Chillu letters are mutually exclusive, the traffic based internet publishing industry, which is in its early days, of Malayalam is largely affected. The same also is affecting the traffic to Malayalam wikipedia potentially reducing the new editors, or even edit counts[5].

Nta Issue

[1] http://malayalam.kerala.gov.in/index.php/InputMethods#Inscript

[2] Malayalam wikipedia - http://ml.wikipedia.org

[3] https://bugzilla.wikimedia.org/show_bug.cgi?id=25623

[4] Michael Kaplan, If the collars and cuffs don't match... http://blogs.msdn.com/b/michkap/archive/2011/07/21/10188466.aspx

[5] Gerard M, The Malayalam Enigma. http://ultimategerardm.blogspot.com/2010/12/malayalam-enigma.html

NA + VIRAMA + RRA - /ന്റ/ is the widely used sequence. This is the sequence as per Inscript , and as taught in Schools of Kerala. This is the sequence followed by all popular transliteration based input tools. And this is the sequence supported by all Malayalam fonts except Microsoft Kartika in latest Windows Operating system.

The NA + VIRAMA + ZWJ + RRA - /ന്‍റ/, which is the supported sequence of Nta for Kartika font is also seen in web since users type in that way to get correctly displayed by Windows default font. Anjali Old Lipi hides this difference of sequences by showing staked nta /ന്റ/ for both sequence.

The Unicode standard for this sequence as explained Unicode 5.1 - Chillu n + VIRAMA + RRA is no where used. No rendering engine, no font or no Operating system supports that.

AU vowel sign issue

Both /◌ൗ/ and /െ◌ൗ/ are popular in Malayalam. Inscript standard uses /െ◌ൗ/ and many other input tools follows that. But there are popular input tools which use /◌ൗ/ for the sign.

Whoever searching with this sign in this word will lose half of the search results.

Potential Future Issues

As mentioned above, widely used characters of Malayalam facing dual encoding issues, the future of Malayalam Unicode is not bright. Standards, which are built on top of Unicode, like ICANN Guidelines for Malayalam is already facing issue[1]. Input method standard revisions are in a dilemma of which encoding to be used[2].

The way forward

This report explained the issues that Malayalam is facing because of the decisions of Unicode in the past. But the report has nothing to propose as a solution. The report intends to bring the attention of UTC into the above mentioned issues and ask discussions and solutions for them. When Unicode introduces the new characters, obviously there is

[1] SMC Critique and discussion on suggested IDN Policy for Malayalam http://wiki.smc.org.in/CDAC-IDN-Critique

[2] Enhanced Inscript standard - Discussions and critique from community - http://wiki.smc.org.in/CDAC-Inscript-Critique

a time delay for implementations to catch up with the latest standard. But when an existing code point or character sequence is redefined or duplicated, more than the implementation delay, the compatibility with the existing data and the data being produced by existing implementations are equally important. In the case of Malayalam Chillu and Nta, Unicode stability policy is violated. But individual implementations cannot break the backward compatibility in that way. That is one of the reasons for “implementation delay” or “not implemented” status of applications. Higher order custom solutions are possible in many places, like the custom conversion rule written in Mediawiki for Malayalam. But such workarounds defeat the purpose of the standard. As far as native language technology developers are concerned, some of them considered the gap between the language and standard and stayed away from the dual encoding implementation and custom conversion solution[1].

[1] Praveen A, Unicode or Malayalam? http://www.j4v4m4n.in/2009/11/07/unicode-or-malayalam/

Appendix 1- Important mails

Dr. Uma Maheswaran, Indic Mailing List of Unicode (8/8/07)

Dr. Ganesan, Ralminov Raminovsky and others ... I have been asked for what I did point to etc. at the WG2 meeting that Michael referenced ... I know this posting of mine is going to bring on a lot of FLAME .. Here I go ..

There have been a number of contributions on the topic of Chillus which finally led the UTC and WG2 to support the encoding of Atomic Chilus for Malayalam. You can also mine this list for all the past discussions/postings with examples etc. on this topic. For me the convincing points were:

a. The samvruthokaram issue .. at the end of a word, we cannot simply assume that a chillu form and chandrakala form are equivalent due to the difference in the meaning of the word.

b. The dual mapping of some .. for example: RA, RRA Chandrakala and RRA Chillu (for me the name Chillu RRA is more appropriate reflecting what I know of its pronounciation to be as RR). If I press the key top marked with a Chillu RRA what sequence should be produced by the keyboard driver / input method (if there is no atomic Chillu RRA)? (Same for LA,TA corresppondence with Chillu LA, and possibly others like that).

c. The example provided about the stacked N-with-R vs side by side N and R in one of the contributions, in my view could not be easily resolved without atomic chillu (at least for) NA chillu.

In my mind, at the heart of the matter are the issues raised by the orthographic reform and its ramifications to the Malayalam writing system, the North vs South Kerala traditions related to Samvruthokaram item mixed up with the orthographic reform etc.. While the Atomic Chillus does raise some backward compatibility issues in the long run, overall it solves more problems than it creates.

The joiners being ignored by TLD IDNs is more recent an issue ... compared to the above.

Michael Everson's posting refers to my pointing to the RA, RRA Chandrakala correspondence with the single RRA Chillu item above at the WG2 meeting which accepted the Atomic Chillus for encoding. Of course, UTC had quite a few discussions before that, even rescinding the decision once, till more arguments could be heard.

I know these are my opinions and there are various arguments put forth ... for and against these.

I also requested that the pro and anti Atomic Chillu groups in Kerala --- experts who are much more knowledgable than myself on the various issues -- get together and arrive at a single consensus position on these issues. I also commented that if a single unified position was coming forth from kerala, we all have to live with the consequences of whatever is the final outcome of the debate, decided by people outside Kerala. Looks like a consensus could not be achieved whatever be the reason.

Finally, the Govt of India's (a full member of UTC) and Govt of Kerala's positions, including a couple of letters from the CM of Kerala (with initially with anti atomic Chillu flavour and later a pro atomic Chillu flavour) were put forward for UTC consideration. Not to mention the Kerala Govt gazette which pointed to an experts group recommending Atomic Chillus about 10 years ago, including a keyboard layout with a few Chillus on them. Arguably the evidences were not documented by that experts group for posterity.

As the recording secretary of the JTC1/SC2/WG2, I took it upon myself to respond to a letter the convener and SC2 secretariat received (I was copied on it) requesing a reversing the Amd. 4 encoding of atomic chillus at the April WG2 meeting, with a pointer to a workshop in Kerala on the topic. I explained the procedure to follow ...i.e. through the P-member body (Bureau of Indian Standards) to respond to SC2 ballots.

I noticed the timing of such letters by the sender(s) .. always sent a couple of days prior to an important meeting and the delegates are traveling; there is iffy internet connections to one's own email beihind some firewall ... and thereby no way to consider these email messages prior to that meeting !!

I also have observed statements like .. 'you do not know anything about my script / language; you dont know what you are talking about'... kind of comments on this list. Fine .. the same commenting experts are not able to convince others who they consider to be 'knowing everything about the script/language' and arrive at a single consensus position !!

In my view, there is no turning back the clock on at least the initial set of Atomic chillus ... and yes, we have to live with that -- good or bad.

Frankly, many of us who are involved in these discussions over the last two or three years, want to go forward with the decision of Atomic Chillus and deal with all its fallouts as we go forward.

Yes, some guidelines about the fallouts have to be crafted and published on backward compatibility issues. Also many of the speculations as to what may or may not be encoded in the future as far as Malayalam and other Indic scripts are concerned, are of concern.

Best regards, Uma
V.S. UMAmaheswaran, Ph.D.
Globalization Centre of Competency, IBM Toronto Lab
A2/SZ8, 8200 Warden Avenue, Markham, ON, Canada, L6G1C7; +1 905 413 3474;
Fax:905 413 4682; TieLine 969; email: umavs@ca.ibm.com

N Ganesan in Indic Mailing list on 8/9/07

Dear K. G. Sulochana,

It is good that we recognise the enormous valuable existing and growing data in Malayalam unicode in the web. In the web, chillus are displayed uniformly now using chillu sequences.

Look at the number of blogs in Malayalam that use chillu sequences and display properly the cils. Backward compatibility is vital here (as Jonathan and Erkki pointed out earlier), and have to be supported indefinitely in the future in rendering engines, fonts, software for Malayalam (cf. Peter's mails in Indic)

Malayalam chillu sequences are not like the case of Myanmar,
but must be supported.

N. Ganesan