L2/06-396

To,
Mark Davis
President
Unicode Consortium

Dear Mark Davis,
Subject: Chillu encoding is wrong

Some days after the intervention of the Chief Minister of Kerala on this matter, some of the responses seen were: "*... Due to the controversies surrounding Malayalam encoding, UTC is moving cautiously ...*" (Rick McGowan), and "*... UTC has since made steps to encode those characters but is intentionally progressing at a slow pace – that way there is momentum to get some resolution to the issues, but time is being allowed for the user community to work out a consensus on the solution...*" (Peter Constable).

Given that the Malayalam encoding debates have lasted this long, it was surprising to see Micheal Everson's passing comment that "*WG2 has accepted to add the six chillu characters to a new ballot for Amendment 4 to ISO/IEC 10646.*" It is from this email, that we even came to know about the WG2 meeting.

It is not proper to present the conclusions regarding such a hotly debated topic, through a tricky passing comment. However, Malayalees cannot accept such a conclusion because such a decision cannot be derived from a proper consideration of facts regarding Malayalam. There has not been a single logical reason or observation to support the proposal for chillu encoding. The document on which the latest decision is based (i.e., L2/06-189) is also filled with fallacies and misleading statements about Malayalam, and technical aspects of Malayalam encoding.

1. First and foremost, the means by which this decision was taken is improper and unacceptable with regard to the Malayalam-speaking community. During the last UTC meeting the Chief Minister of Kerala, Mr V.S. Achuthanandan has officially intervened in the prolonged disputes over Malayalam encoding, and requested the Unicode Consortium to stop the encoding process until a consensus could be formulated.
2. Democratically speaking, Mr. V.S. Achuthanandan is the duly elected representative of Malayalees. He is the premier authority as far as the affairs of the Malayalee community is concerned. He is also the authority to express the aspirations of the Malayalees.
   As a response to his request, the present action to proceed with the encoding is inappropriate.
   In fact, this jeopardizes the formation of the committee with eminent scholars, computer scientists and IT experts which was ongoing since the Chief Minister sent his message to the UTC.
3. Since the decision is based on L2/06-189, it is further questionable, because, this document is thoroughly wrong, factually and historically, it misinterprets, and tries to mislead others on the linguistic and technical aspects of Malayalam encoding.
   Even persons with very limited knowledge of the history of Malayalam language are quite appalled at the observations made in that document. The set of examples given as Malayalam by the authors on page 1 of the document is especially disgraceful; they are contrived, unnatural constructions intended to mislead the reader who is not familiar with the language; they are even insulting to the Malayalee people.
4. The consideration given to the persons who submit conjecture and speculation as "evidence" on the mailing lists and directly to the UTC without resort to facts is one of the failures of the process instituted by UTC. That many of these documents are taken as expert opinion and referenced as statements of fact is appalling.

For instance, “Malayalam” examples like മന്വിക്ഷോഭം/മൻവിക്ഷോഭം, കണ്വലയം/കൺവലയം which disregards even the basic word formation norms of Malayalam. Actually, very few Malayalees access the Unicode mailing lists which contains these fantastic samples; these examples have only caused extreme laughter when we show it to them.

5. We reiterate that this move is totally against the character model of Unicode, which is to encode only basic characters. It is very common in most languages in the Unicode, for characters and sequences to have multiple graphical manifestations. The chillus are no different from those.
   The chillus fall very clearly into the Indic encoding model and they can be represented very easily using the facilities developed for Devanagari and other Indic languages. There is absolutely no need for separate codepoints for chillu, since there is already a method to encode chillu forms which will continue despite the new codepoints. If this decision is carried out, then there arises a dual encoding for the chillu and thus a serious threat of security issues.

The UTC has the ultimate responsibility to validate the linguistic sanctity of such documents which makes fun at a language, its logical structure and its users.

In fact, our arguments from the very beginning have always been supportive of the original stand of the UTC with regard to Malayalam/Indic encoding, and we are not trying to bring something new. We do not understand why the UTC is taking a stand against its original standards which is logical, practical and correct. In conclusion, we strongly attest that this chillu encoding will lead to a dual encoding and attendant problems. This will lead to huge confusion in the use of Malayalam for computing, perhaps worse than what exists today.

Analysis of L2/06-189

This document is full of factual errors and cannot be used as the basis for any decision on the chillu issue.

The document starts with an error by categorically stating:

> “They are unique characters in Malayalam, which cannot be represented by existing
> characters in the Malayalam Unicode code space (0D00 – 0D7F).”

Firstly, the chillu characters are not unique characters, since they are presentation forms of other basic characters of Malayalam. This has been shown time and again.

Secondly, there is ample evidence to show that each chillu is derived from only one base character. Statements in other documents submitted to the UTC, in which chillu are shown to arise from more than one base character is conjecture and/or error in interpretation of the referenced texts.

For e.g., ർ arises only from ര and not from റ. In many words, the pronunciation of ർ is similar to റ with vowel ommitted. However, this is due to the regular phonological features of Malayalam.

The second paragraph,

> “There is difference in meaning, when chillu consonant + chandrakkala combination is
> used in place of Chllu.”

is deemed to show a constrastive use of the chandrakkala and the chillu. We have already discussed this issue in detail in our submissions to UTC.

All the examples above show the chandrakkala as a manifestation of the psuedo-samvruthokaram. What is missing is that chandrakkala can also manifest vowellessness (in consonants with overt-chandrakkala), and chillu is merely vowellessness of their respective base consonants.

The chillu encoding is supposed to distinguish between the samvruthokaram and vowellessness. However, it cannot be successful since the chillu encoding cannot distinguish between all instances of samvruthokaram, psuedo-samvruthokaram, chillu, and consonants with overt-chandrakkala.

The next paragraph is,

> “Use of joiners for linguistic functions will result in overloading the joiners/using them
> for purposes other than their defined functions.”

The use of ZWJ and ZWNJ to produce the chillu and overt-chandrakkala forms is already defined in the Unicode Standard. Thus, they are not “overloaded”.
Linguistically, the functions of the ZWJ/ZWNJ in Malayalam mirror those of Devanagari. The ZWJ/ZWNJ is harmonious with the inner grammar of Indic scripts, and their encoding in Unicode.

It should also be noted that if ZWJ should be removed from Indic encoding by whatever argument, the very same argument should also be applied to the ZWNJ. Of course, removal of ZWNJ would entail either the creation of a new codepoint for a non-joining chandrakkala, or require encoding of the vast number of conjuncts, which are presentation forms of sequences of consonants.

Both ZWJ and ZWNJ are integral parts of the Indic encoding model, and their properties match well with the requirements of Indic.

The last point made is,

> “Last and the most important, As per the IDN standards, the format characters like
> ZWJ, ZWNJ etc are prohibited in domain names. Because of this with the present
> representation of chillus using ZWJ, we will not be able to use a large number of
> words like സർകാർ, തൊഴിൽ, സിവിൽ etc. in domain names.”

It is a fallacy to assume that because ZWJ/ZWNJ are prohibited in ACE domain names, they should be impossible to use in user interfaces.

The reality is that ZWJ/ZWNJ are prohibited *to have a mapping to anything but the null string* and not that they are prohibited per se when entering the domain name in the URL text widget of a browser. Since the ZWJ/ZWNJ are mapped to the null string (see Table B.1 and C.2.2 of RFC 3454 StringPrep), they do not reach the “Prohibit” step of Stringprep preparation of strings. Also, mapping them to null string, is advantageous to Indic encoding in that presentation variants are mapped to the same Punycode string.

```
Python 2.4.3 (#2, Apr 27 2006, 14:43:58)
[GCC 4.0.3 (Ubuntu 4.0.3-1ubuntu5)] on linux2
Type "help", "copyright", "credits" or "license" for more information.
>>> from encodings.idna import *
>>> ToASCII(u'സര്‍കാര്‍') #Both ര + ◌് + ZWJ
'xn--bwc7cb3a6a2fd'
>>> ToASCII(u'സര്കാര്◌') #Both ര + ◌്
'xn--bwc7cb3a6a2fd'
>>> ToASCII(u'സര്കാര്‌') #Both ര + ◌് + ZWNJ
'xn--bwc7cb3a6a2fd'
>>>
```

This also has security considerations: It is advisable for applications that render text, to provide for shaping a sequence C_1 + chandrakkala + ZWJ to a chillu glyph, even if the chillu characters are encoded, to preserve backward compatibility. However, this leads to a dual encoding, since now there would be 2 ways to represent a chillu.

One way to solve this problem, would be to have language specific nameprep tables for Malayalam, which map the chillu characters to their base forms. Doing that however, would invalidate the chillu codepoints since that would render the first argument of L2/06-189 useless, i.e., that chillus are “unique characters”. For e.g., ൻ would be mapped to ന + ◌്. In the case of ൽ, however, which base consonant is to be chosen ? The proponents of the chillu codepoint theory would have that ൽ is derived from both ല and ത and in this case, makes the security issues even worse.

The correct interpretation is that each chillu is a presentation variant of the vowelless forms of unique consonants, i.e.,

ൺ	ണ + ◌്
ൻ	ന + ◌്
ർ	ര + ◌്
ൽ	ല + ◌്
ൾ	ള + ◌്

If such a mapping is not made, then the number of possible spoofed URLs increase exponentially with the number of chillus in the domain name.

സിവിൽ	സിവിൽ
	സിവിൽ
സർക്കാർ	സർക്കാർ
	സർക്കാർ
	സർക്കാർ
	സർക്കാർ

Table 1: Red is for chillu codepoint and Blue is for chillu represented using base consonant + chandrakkala + ZWJ

As mentioned in our earlier documents submitted to UTC, it is a feature of Malayalam that a single character may have multiple manifestation; this is not at all accidental, but are due to many historical reasons.

The authors also attach the note from the .in authority on their mapping table submission:

> “The table does not conform to RFC 3491 (Nameprep). Some of the characters should not be in the table (eg, 200C, 200D - control characters) which are explicitly banned in the DNS. We request that the team at C-DAC review RFC 3491 and ensure that the table conform to this RFC.”

In this regard, we too would like to ask the authorized person at CDAC to review RFC 3491 and other relevant documents on IDN before making mapping tables. It is entirely irresponsible to publish a note from the .in authority as if it were a recommendation against ZWJ/ZWNJ, when in fact the fault lies with the authors who tried to add ZWJ/ZWNJ into the mapping table which is clearly prohibited. We also recommend the authors of L2/06-189 to review their understanding of the IDN model before making any comments on the same.

We also request CDAC to use a public process to decide whether mapping tables are indeed required in this case (we believe it is not necessary in Malayalam), and if so, what the table would be like.
By this we are not making a comment on CDAC as a whole: we know several people within CDAC who do not share the same viewpoints regarding Malayalam encoding as that of the authors of L2/06-189, and in this particular case, it is the individual prejudice and ignorance of language, language computing and relevant standards, of the authors that caused the .in authority to retort on errors in their submission.

On page 2, the authors provide the following presentation:

The authors reproduce our 3-point solution for the chillu issue:

> “According to Rachana, the solution for solving the Chillu issue is as follows.
>
> 1. Accepting only ( ) (0D41 + 0D4D) form as the samvruthokaram, following its predominance in the original system,
> 2. Giving the chandrakkala the sole function of minus-vowel marker when used with consonants, and
> 3. Retaining the existing situation of chillaksharam using Joiners”

When we stated our 3-point solution to the samvruthokaram, we were not prescribing the use of samvruthokaram, or what a user may or may not do. These were provided as design guidelines for computer programs: while the user may use psuedo-samvruthokaram forms in their writing for e.g., writing അത് (psuedo-samvruthokaram) vs. അതു് (samvruthokaram), computer programs cannot distinguish in a given word whether it is in fact a pseudo-samvruthokaram or overt-vowellessness (e.g., in തത്, it could be either pseudo-samvruthokaram or overt-vowellessness). In this case, the program must use assume that the word represents overt-chandrakkala since that is harmonious with the logic of Malayalam script.

In higher level programs where more information is available, this 3-point solution may be tailored e.g. in grammar checkers, the tagger would know that തത് is a Sanskrit word and thus the chandrakkala at the end is indicative of overt-vowellessness and not of psuedo-samvruthokaram. Similiarly, the same tagger, would also understand that കാത് (which means ear) is a Malayalam word and thus the chandrakkala at the end is indicative of psuedo-samvruthokaram and not overt-vowellessness.

In lower level software where such a determination is not possible, it is only logical to use the 3-point solution.

In page 2 of the L2/06-189, regarding the samvruthokaram; it is a deliberate attempt to mislead UTC on the history of samvruthokaram.

The vast majority of printed materials available today uses the samvruthokaram form rather than psuedo-samvruthokaram. It was only after the introduction of Typewriter keyboard into DTP in recent times that newspapers were forced to avoid samvruthokaram.

The most authentic Malayalam dictionary, Sabdatharaavali referred by millions everyday, uses samvruthokaram. If UTC requires it, we will provide additional proof that vast majority of printed materials use samvruthokaram.

In light of the above facts, we request that UTC not take hasty decisions based on illogical documents, which gives false information and constructed examples specifically in order to mislead the understanding of Malayalam. These observations do not represent the views of Malayalam scholarly public. We uphold the original Unicode standards without any amendments in this regard. The dual encoding of chillus cause great harm to our language. The added confusion will lead to broken implementations and differing viewpoints regarding the use of codepoints. Additionally, this alienates Malayalam from the general structure of Indic scripts designed by Unicode.

We reiterate that the UTC must give due consideration to the message of the highest elected representative of Malayalees, Honbl. Chief Minister of Kerala, on behalf of the Govt. of Kerala and of the People, and not to take a hasty decision until a consensus opinion can be reached regarding Malayalam encoding including chillu issues.

Regards,

R. Chitrajakumar,
N. Gangadharan,
Rajeev J Sebastian
Rachana Akshara Vedi