L2/05-308

Problems of Malayalam Encoding in the Indic context

(Rachana's response to Malayalam encoding debate)

R. Chitrajakumar and N. Gangadharan

Rachana Akshara Vedi

1 ന്റ /n̲t̲a/ and ൻറ /n̲r̲a/

The confusion surrounding the issue of ന്റ /n̲t̲a/ and ൻറ /n̲r̲a/, is one of the main issues that are raised in order to encode the chillus.

Even though we had precisely described the ന്റ /n̲t̲a/ and ൻറ /n̲r̲a/, in section 6 of our earlier Rachana (L05-210) document, some persons had expressed certain confusions over this issue. Some commented that our description was too brief. Below we provide a more detailed treatment of this issue. Please read along with the description we had provided in L05-210.

The ambiguity and confusion regarding those sequences is due to the lack of specific basic characters for the alveolar stop (റ /t̲a/) and alveolar nasal (ന /n̲a/).

1. In all Indic languages, there are 5 consonant classes (vargas) velar, palatal, retroflex, dental and bilabial. But, in the case of Dravidian languages, there is an additional alveolar class. This class is used with high frequency in Tamil and Malayalam. It is used in a large set of words of the core vocabulary.
2. Another feature to be noted in this context is that in Indo-Aryan languages each class (varga) has 5 instances, whereas in Dravidian languages, there are only 2 (stop and nasal).
3. Unlike Indo-Aryan languages which have many more possible combinations, both within the class (varga) and also between classes, in Dravidian languages, the number of combinations are strictly limited. Each class produces only 3 combinations: 2 geminations and 1 conjunct with the nasal as first member. Also, inter-varga combinations do not exist[1]. For e.g., in the velar class the three conjuncts are
   1. gemination: ക്ക /kka/, ങ്ങ /ṅṅa/
   2. combination: ങ്ക /ṅka/
4. As shown in chart 3, row 6, like all other classes, the alveolar class has the same feature and behavior, i.e., basically it has a stop and a nasal, and combinations are റ്റ /t̲t̲a/, ന്ന /n̲n̲a/ and ന്റ /n̲t̲a/.
5. In Malayalam റ /t̲a/ and ന /n̲a/ do not have their own glyph. They are attained using the glyphs of ർ and ന. This is due to some historical reasons (also, grammatically justifiable). In Tamil, ന na has a seperate character from ന /n̲a/, and ற /t̲a/ is attained using the glyph of ர /r̲a/. Thus, the usage of ற for /t̲a/ came from Tamil directly. In the case of ന /n̲a/, in Malayalam, the glyph of

1 The conjuncts ന്മ /n̲ma/ (alveolar ന /n̲a/ and bilabial മ /ma/) and ണ്മ /ṇma/ (retroflex ണ /ṇa/ and bilabial മ /ma/), are not natural conjuncts. They are the resultant form of some phonological changes.

dental ന /na/ is used for it, because it can be used unambiguously.

6. In Malayalam, for the sound /ṯa/ there is no independent glyph. So, it is written as റ as in the case of Tamil. Both /ṯa/ and /ṯṯa/ have the same glyph, i.e., റ. When it appears in natural Dravidian conjunct ന്റ /n̲ṯa/, only one റ is used[2].

The circumstance that all vargas and combinations have distinct glyphs, and that only the characters and combinations of the alveolar varga uses the glyphs of other characters, is the root cause of the confusion regarding the ന്റ /n̲ṯa/ and ൻറ /n̲ṟa/.

ക ka	ഖ kha	ഗ ga	ഘ gha	ങ ṅa
ച ca	ഛ cha	ജ ja	ഝ jha	ഞ ña
ട ṭa	ഠ ṭha	ഡ ḍa	ഢ ḍha	ണ ṇa
ത ta	ഥ tha	ദ da	ധ dha	ന na
പ pa	ഫ pha	ബ ba	ഭ bha	മ ma

Chart 1: Indo-Aryan Classes (Vargas)

ക ka	ങ ṅa
ച ca	ഞ ña
ട ṭa	ണ ṇa
ത ta	ന na
പ pa	മ ma
റ ṯa	ന n̲a

Chart 2: Dravidian Classes (Vargas)

1	ക ka	ങ ṅa	=>	ക്ക kka	ങ്ങ ṅṅa	ങ്ക ṅka
2	ച ca	ഞ ña	=>	ച്ച cca	ഞ്ഞ ññ a	ഞ്ച ñca
3	ട ṭa	ണ ṇa	=>	ട്ട ṭṭa	ണ്ണ ṇṇa	ണ്ട ṇṭa
4	ത ta	ന na	=>	ത്ത tta	ന്ന nna	ന്ത nta
5	പ pa	മ ma	=>	പ്പ ppa	മ്മ mma	മ്പ mpa
6	റ ṯa	ന n̲a	=>	റ്റ ṯṯa	ന്ന n̲n̲a	ന്റ n̲ṯa

Chart 3: Dravidian Combinations

We have seen that, ന്റ /n̲ṯa/ is a conjunct of vowelless alveolar nasal and alveolar stop, whereas in ൻറ /n̲ṟa/, it is a sequence (not a conjunct) of vowelless alveolar nasal, followed by റ /ṟa/. It should be noted that both sequences begin with ന, just as in pronunciation and writing.

When considering the conjunct making mechanism as used by Malayalees,

1. ന്റ /n̲ṯa/ must be inputed as ന followed by റ as a conjunct
2. ൻറ /n̲ṟa/ must be inputted as ന followed by റ but not as a conjunct.

This is the mental model of the user and it appears in exactly the same way in rendering and encoding.

The total process that occurs here is that in one sequence റ forms a conjunct with ന and in the other, it follows the ന without forming conjunct. In both cases,

1. the process begins with the basic element ന
2. then an operator is added which is used to select whether the following റ forms a conjunct with the preceding ന or not
3. finally, the sequence is completed by appending റ

2 The same letters are also used in alveolar loan contexts.

When considering the chillu-na codepoint used for the purpose of disambiguating ന്റ /n̲t̲a/ and ൻറ /n̲r̲a/, this model is broken and the chillu codepoint which is equally applicable to both sequences, will only be used for one.

When proposing a mechanism for inputing the chillu na codepoint, there will be a key on the keyboard to produce it on the screen. When this key is present, both ന്റ /n̲t̲a/ and ൻറ will be inputted with the same key, i.e., both sequences will be inputted as <ൻ chillu-na + റ r̲a>. This makes the disambiguation of ന്റ and ൻറ via chillu codepoint a failure.

This so-called problem of ന്റ /n̲t̲a/ vs ൻറ /n̲r̲a/ has nothing to do with ന. Rather, it is a problem of റ and whether it forms a conjunct with the preceding ന. Here the very source of confusion is the conjoining manner of റ with ന in one case, and the non-conjoining manner of റ with ന in the other case, i.e., the problem is not with ന, rather it is with റ. When the fundamental problem here is about conjoining and non-conjoining റ, suggesting codepoint for chillu-na as a solution is quite unreasonable.

In this context, we can see that trying to solve the problem by encoding chillu is not at all feasible and in fact increases the confusion and violates the mental process which is same in writing.

2 Graphemes and Encoding

In an encoding, a letter or grapheme of a language is manifested in a very mechanical manner just as in established writing, i.e., we encode accepted written symbols of the language. It is very important to note that, the phonological and underlying features of a particular letter or grapheme should not be overloaded in the consideration and encoding of that letter.

For example, the English 'n' is used for alveolar, dental and palatal sounds. At encoding level, no weight or consideration is given to this fact for encoding the 'n'.

In a similar fashion, ന is used for both alveolar and dental nasals. The vowelless manifestation of this letter as chillu has the value of vowelless ന and is a presentation variant of vowelless ന. Through ZWJ/ZWNJ mechanism, chillu is produced from the ന. Thus, the derivation of the chillu form is clear and unambiguous.

When this chillu-na appears in a conjoined environment and a non-conjoined environment, it behaves as vowelless ന without any confusion. Over-interpreting the appearance of chillu-na is a meaningless pursuit.

Similiarly, as we stated above, in all contexts of conjoining and non-conjoining sequences of ന്, ൻ, etc the ന is the underlying consonant and it can appear in the following forms:

1. vowelless consonant
2. chillu
3. chillu followed by another consonant which has value as a natural Dravidian conjunct (ന്റ /n̲t̲a/)
4. chillu written separately followed by a consonant, but does not have a value as a conjunct (ൻറ /n̲r̲a/)
5. probable common conjuncts of ന with other consonants (ന്മ /n̲ma/)

The mystery behind പിൻനിലാവും /pin̲nilāvum/ and പിന്നിലാവും /pin̲n̲ilāvum/, കണ്വലയം /kaṇvalayam/,

etc is because of the non-consideration of the above mentioned facts. In all such examples, vowelless consonant and its manifestations as chillu, conjuncts, etc appear. It is not at all a matter of confusion as far as encoding and rendering are concerned. Also, for higher level applications, the use of ZWJ/ZWNJ is correct.

3 Perspective of Malayalam computation

Initially, some persons requested separate chillu codepoints. After browsing some grammar books, they have discovered some additional characters which they claim to be chillu. In fact, they have even requested the encoding of യ-chillu which the Malayalee has never even imagined.
Continuing in this fashion, now they are asking for codepoints for the consonant marks like post base യ, വ, etc, placing examples of താഴ്വാ /tāḻvō/, etc. They have gone from chillu encoding, to requiring many more codepoints just for presentation variants.

Furthermore, the confusion regarding the fundamental value of chillu leads these persons to come to the fantastic conclusion that the post base marks of the chillu forms, already requested for encoding, should also be encoded as codepoints. Thus, for e.g., codepoints would be necessary for ര /ra/ and ല /la/, ർ /r/ and ൽ /l/, and their C_2-conjoining forms, as they have explained in താഴ്വാ /tāḻvō/, പന്ത്രണ്ടു /paṉtraṇṭu/, etc.[3]

In effect, what is clear is that such views are based on a flawed understanding of encoding and what constitutes a codepoint according to Unicode standards. The interpretation of a codepoint or a sequence of codepoints may be dependent on the particular application, and also surrounding characters, as mentioned in various Unicode technical annexes and reports. The requests to encode chillus and other rendering variants is either a misinterpretation of Unicode standards, or a misunderstanding of their purpose.

The present debate conducted by UTC regarding Malayalam chillu is to foster an academic discourse on representing it within the Unicode framework with consideration to:

1. it should satisfy the Unicode standards
2. it should in compliance with technologies surrounding Unicode
3. it should be compatible with the general linguistic principles
4. it should conform to the inner logic of Malayalam
5. it should be suitable for Malayalam language applications at all levels without contradictions
6. it should make adequate provisions for future application

In short, the perspective that should be taken is that the encoding is being developed for use. Rachana Akshara Vedi has a clear and firm view on the present Unicode Malayalam embedding discussion.

On this eve of laying the foundation of Malayalam embedding for International use, we are stressing the real features and problems that affects Malayalam computing and applications at various levels. In this context, it is also our special concern to point out the inconsistencies that emerged through the Typewriter reform and other instances, which create problems for Malayalam computation.

We do not wish to put forward linguistic/computational theories. Instead, our aim is to work out the problems of Malayalam, well within the principles, standards and framework of Unicode and to realize the full potential of Unicode technology for Malayalam computing and applications.

3 http://varamozhi.blogspot.com/2005/07/unicode-issues-with-visible-virama.html

For this purpose, without any prejudice, we thoroughly went through the claimed problems of the encoding, the problems that have occurred to us, and also of the new proposals. We have the responsibility primarily towards Malayalees, and also to the scholarly public knowing the depth of Malayalam, who are our well-wishers and who constitute Rachana Akshara Vedi.

Unfortunately, the request for chillu codepoints seems to be highly prejudiced without considering the above factors. The manner, mode of presentation and the spirit of the arguments are such that - some persons have almost assigned the chillu codepoints and so Unicode and Malayalees should accept those desires for whatever reason.

All the argument statements and the total spirit of the arguments reflects this mentality of keeping the cart before the horse.

It is very difficult to accept such a viewpoint as far as a live language like Malayalam is concerned, which has a logically and grammatically rich script.

4 Original vs Typewriter in the context of Unicode

When considering some of the statements and technical discussions with regard to embedding Malayalam in Unicode, one gets the impression that there are two entirely different script systems in use in Malayalam, i.e., the Original and Typewriter script. Some persons build up entire arguments on the basis of this claim. However, this viewpoint is not correct.

Some comments regarding the Unicode encoding such as, "It should be an inclusive design for both old and new orthographies. We can not simply reject one and encode only the other." , etc, is indicative of a general trend in the Unicode Malayalam community.

There is also a contention that the arguments put forward by Rachana Akshara Vedi regarding Malayalam script in Unicode, is directed towards and supports only the Original script.

In the 70's, the said reform was implemented, with the main objective of bringing Malayalam script, consisting of vowels, consonants and conjuncts, to the Typewriter keyboard. The reforms were the following:

1. The vowel marks of ു /u/, ൂ /ū/, ൃ /ṛ/ were detached from the consonants
2. The ര /ra/, which originally formed conjuncts with consonants, was also detached and instead, a prebase mark was introduced
3. All conjuncts were split and written as sequences of basic characters with chandrakkala
4. Samvruthokaram was replaced by the chandrakkala (pseudo-samvruthokaram)

However, the Reforms committee specifically directed that these reforms were not to be applied in writing or in education, and that it was intended purely for the Typewriter. But, it appeared in some Government level printing and slowly spread into the writing of a minority.

When this Typewriter script appeared in typed sheets and printing, it was heavily criticized. Leading

publishers and media persons as a whole denounced this reform and kept away. The Malayalam script that was systematic and standardized without any issues, now became inconsistent.

When Typewriter keyboard began to be used in DTP systems, this script started to appear on the computer. The majority of people were not able to fully avoid it, and were forced to accept it to a certain extent.

Actually, there is not much fundamental difference between the Typewriter script and Original script. As we have seen earlier, the reforms occurred only in a very small area of the Malayalam script. As far as basic characters of the script are concerned, no changes have occurred at all.

All changes occurred only in the case of three vowel signs and one consonant sign which were detached and instead some marks were introduced. The splitting of conjuncts with chandrakkala has not added any basic characters or changed the formation of syllables: the result of splitting conjuncts are the pre-existing basic characters of Malayalam.
Neither changes in the vowel signs, nor splitting of conjuncts, affects the basic features of characters. In short, all the changes occurred only in glyphs and not in the basic characters.

Even though there were no changes in the basic features of the script, the splitting of conjuncts lead to great confusion. From the beginning of the reform, Malayalees could not accept the splitting of the conjuncts. The splitting disturbed the inner logic of conjunct formation, and also the aesthetics of the script. Due to this, from the beginning, there were always moves back towards the Original script.

There were reforms to Typewriters themselves, with new designs that could type more conjuncts. In spite of the 256-glyph limitation, more and more software packages began to use more conjuncts. There has been a steady return towards the Original script. The Typewriter reform has contributed only a thorough inconsistency in the well organized script of Malayalam.

All problems caused by the Typewriter script emerges at higher levels of usage.

In conclusion, there is no basic difference between the Original and Typewriter scripts as far as encoding is concerned. In encoding logics, there is no difference between how the Original and Typewriter scripts are encoded because both are represented using the same basic characters: the possibility of viewing the same Malayalam Unicode text with both Original script and Typewriter script fonts is a testament to this generic encoding.

Thus, the contentions and observations that some arguments are applicable only to Original script and not applicable to Typewriter script, and that Unicode should give equal importance to both Original and Typewriter script, creates an impression that there are two completely different scripts and is therefore baseless.

The only feature to be commented is that of the samvruthokaram. The samvruthokaram is not a basic character: it is a sequence. At encoding level, the replacement of the samvruthokaram with the chandrakkala does not create any problems, i.e., there is no necessity for any new characters, and the chandrakkala exists in both the Original and Typewriter scripts. This is true not only in the case of samvruthokaram, but nowhere in the entire context of Typewriter script is there a need for new characters separate from the Original script.

Even though there are very grave and complex problems created by Typewriter script at higher level, at the

encoding level there is no such difference to be considered between the Typewriter and Original scripts.

5 Devanagari Perspective

As we have stated earlier, chillu expresses the vowellessness of certain consonants. While the usage of vowelless feature in Devanagari, etc is high, in Malayalam, it is certainly limited to some letters, which is a Dravidian feature. Please refer our document “Chillaksharam of Malayalam Language”.

The chillu are manifestations of the above said vowelless consonants. In Devanagari, there are two schemes for denoting vowellessness.

The first case is that, in Sanskrit, vowellessness is denoted with a visible halant. In Hindi, etc, even though there is a special mark (halant), even without this mark, consonants in word-ending position are pronounced and considered as vowelless. The explicit halant rendering of Sanskrit is visually and value-wise the same as the chillu: in Sanskrit the 'tail' goes downwards, whereas in Malayalam it goes upwards!

Secondly, in Sanskrit, Hindi and modern Indo-Aryan languages, consonant clusters is not a random feature, but a strong, regular and basic feature. The formation of conjuncts in consonant clusters has definite rules also. It is these very rules that are also applicable in the case of Malayalam: these conjuncts are purely for expressing the conjuncts in loan words, mainly from Sanskrit. For pure Malayalam conjuncts we described in section 1 of this document. The scheme and particular glyphs and marks are the only differences between the conjuncts of Devanagari and those of Malayalam.

However, in the two features mentioned above, i.e., chillu (vowellessness) and conjuncts, are equally present in both Devanagari and Malayalam.

In Unicode, the mechanism for controlling and restricting conjunct formation is equally applicable to both Devanagari and Malayalam. It is successfully achieved and totally settles all problems using ZWJ/ZWNJ.

Our strong conclusion in this regard is that, what is successfully implemented in Devanagari should not be dragged into debates in Malayalam.

6 Conclusion

1. Chillu forms should not be given seperate codepoints. It is unscientific and illogical according to the basic norms of the Malayalam language and encoding principles of Unicode.
2. No advantages have been pointed out for the chillu encoding so far in the discussion. On the other hand, the chillu codepoints violate a fundamental principle of the Unicode by introducing 2 basic characters to represent the same value. In all cases, chillu is only one among the manifestations of vowellessness. It is not an alien feature of Malayalam, rather it is a common feature of all Indic scripts, to have different manifestations or renderings for the same characters and conjuncts. The context of writing conjuncts in split form is also quite common.
3. Similiarly, vowellessness is not isolated to chillu; similar concepts also exist for Devanagari as half-forms. The explicit halant rendering of Sanskrit is visually and value-wise the same as the chillu.

4. We firmly observe that the confusion regarding the chillu encoding is entirely the lack of clarity regarding the ZWJ/ZWNJ and its application in the Indic encoding model.
   ZWJ/ZWNJ devices of the Unicode technology is logical and sufficient to achieve all the above mentioned contexts. In the case of Devanagari, to achieve the same, ZWJ/ZWNJ is successfully implemented without any confusion.
   ZWJ/ZWNJ is a blessing, as in the case of Devanagari, etc, and it conforms to the linguistic and computational expectations of users.
5. It is not necessary to re-open an issue in Malayalam, which was already settled in Devanagari.
6. Rachana fully supports and upholds the basic character set and the principles of Unicode laid down to formulate it, considering Malayalam as a member of the Indic family.
7. It is high time for Malayalam to lay a strong foundation for different computational applications, compared to other developed languages.