Report of the Workshop on

“Problems of Unicode encoding of Malayalam”

conducted at University of Kerala

on January 24-25, 2007,

organized by the

International Centre for Kerala Studies

Table of Contents

On January 24-25, 2007, the International Centre for Kerala Studies, in association with Dept. of Linguistics, Oriental Research Institute and Manuscript Library, Dept. of Malayalam Lexicon and Dept. of Malayalam, organized a workshop on Unicode Malayalam encoding. Language scholars from these departments and IT experts took part in this event, held at the Senate Chamber of the University.

Prominent IT organizations such as Kerala IT Mission, CDAC, CDIT, NIC, etc were invited to participate in this workshop. Kerala University Vice Chancellor, Dr. M.K. Ramachandran Nair inaugurated the workshop. Dr. D. Benjamin, Dean of Oriental faculty, Univ. of Kerala presided over the function. The veteren scholar of Malayalam, Prof. Panmana Ramachandran Nair delivered the keynote address.

Dr. N. Sam, Director, International Centre for Kerala Studies, gave a general introduction to the problems of Malayalam Unicode encoding. Prof K. Sashidharan (Professor, Govt Engg. College and Director, Information Systems, Finance Dept. of the Govt of Kerala) and Rajkumar (Linuxense) introduced the issues of the encoding from the technical perspective, and outlined the areas where consensus was needed. These introductory talks gave a definite direction to the ensuing discussions. Dr. E.V.N. Namboothiri, Professor of Linguistics, Prof. Dr. P. Venugoplan, Editor, Malayalam Lexicon also spoke on the occassion.

Following this, the delegates divided into two groups to carryout thorough discussions on the linguistic and technical issues. The norms for the discussions were the following,

1. Disputed areas were to be enumerated and discussed one by one
2. The points raised by Govt of Kerala Unicode encoding committee, CDAC, Rachana Aksharavedi, Indic mailing list particpants, Varamozhi group, etc were to be taken/ one by one for discussion.
3. Discuss and finalise the various norms for preparing a comprehensive character repertoire, and to prepare a comprehensive list of historically justified and linguistically valid glyphs comprising of vowels, consonants, signs and conjuncts.

Topics discussed in Group 1

1. Issues related to various aspects of chillu encoding
2. IDN issues
3. Collation
4. Encoding of disputed glyphs like ന്റ, ൺഠ etc

5. Chillu polyvalency
6. ZWJ/ZWNJ issues
7. Input methods

Topics discussed in Group 2

1. Fallback rendering
2. Theoretical basis of character set
3. Malayalam character repertoire

1 Chillu Encoding

Chillus are pure forms of consonants, and they can occur alone without vowel at word-end position.

Due to many historical reasons, the chillus used only in word-end position began to be used also as first member of conjuncts, and also in compound words in word medial position.

In all the above positions it has the value of pure consonant (i.e., without vowel). This is evident from ൻമ, ൽപ: they may be split only as ന്മ and ല്പ.

നന്മ, നൻമ and നന്മ all have the same value. Even though ന in നന്മ, നൻമ and നന്മ differ in rendering, all have same value. Here the question is whether it is right that this ന should be given multiple encodings.

In the current Unicode standard, the ZWJ is used to manifest the chillus.

On the basis of this primary information regarding chillu, the meeting examined the different questions surrounding the chillu issue.

2 IDN Issues

1 Background Information

In IDN registration, the problems caused by chillu encoding are the most widely discussed recently.

As in document L2/06-189, ZWJ/ZWNJ the formatting control characters, are not allowed in domain names for e.g., സർക്കാർ, തൊഴിൽ, സിവിൽ. Thus words with chillu cannot be registered. So, the chillu encoding could solve this issue.

Other argument regarding this is the contrast between വന്യവനിക/വൻയവനിക and മനഃക്ഷോഭം/മൻവിക്ഷോഭം, etc due to which only one in each pair may be registered, and the other folds to the registered one.

It was also noted that the semantically same sequences നന്മ and നൻമ could be held by two websites at the same time.

2 Observations

Workshop delegates observed that these examples are constructed and quite unnatural as far as Malayalam word formation rules are concerned.

Even though they are constructed, the originators of this argument should have approached this issue from another angle: since these 2 constructions have the same underlying sequence, they should be considered the same even though they have different renderings, irrespective of whether they are meaningless or valid words in Malayalam language.

Linguistically, the sequence ൻയ and ന്യ have the same logical constituents: ന + ് + യ.

In practice, such sequences do not occur in Malayalam; the examples provided for study are constructions which do not appear in any Malayalam dictionary and also fails the Malayalam word formation rules. Whether

in this example, or any such examples that may occur in any context, we must accept this as a natural feature of language use in any language, consider the examples as logical sequences rather than as words and expect the two overlapping sequences to fold into a general case. Also, if a desired domain name has already been registered and used (either with the same rendering or with a different one), the registrant has a variety of options such as use of hyphens, underscores, etc. This is a common occurrence on the Internet today.

So, in such cases of IDN registrations with high security implications, it is imperative that an automatic folding mechanism is available for such common situations of similar and ambiguous sequences, so that if one is registered the other should not be available for registration.

On the other hand, if chillus are also included, it will be possible to have multiple representations for sequences that have exactly the same linguistic value, i.e., due to stability policy and backwards compatibility concerns of rendering program, it will be possible to register a domain name having same rendering as one which has already been registered using the ZWJ and the new chillu codepoints. In the best case, it is confusing for users and in the worst case, it leads to spoofing.

The delegates also cautioned the UTC that spoofing causes US $ 2 billion worth of losses to businesses in the USA every year. In the case of Malayalam, spoofing involving chillus does not have to resort to fonts, but rather only to perfectly working rendering engines.

For e.g., the domain name സർക്കാർ.com can be represented in 4 different ways using a combination of ZWJ and new chillu codepoints.

The spoofing issue is also applicable to നന്മ,നൻമ and നന്‍മ, i.e., different manifestations of the same sequence ന + ് + മ must have the same value.

The workshop noted that using mapping tables to solve the spoofing issue is counter-productive since the very reasons for encoding chillu are invalidated, for e.g., it will no longer be possible to disambiguate ൻയ and ന്യ.

3 Recommendations

1. Do not use chillu codepoints for IDNs
2. Do not give any value for ZWJ/ZWNJ in domain names and reject PRI-96

3 Representation of nta/nra

1 Background Information

ന്റ has the pronounciation /nta/ and ൻറ has the pronunciation /nra/. There is an ambiguity representing the nta and nra as in എന്റെ and ഐൻറി. It is suggested that chillu encoding is a remedy for this.

2 Observations

Firstly, the ൻറ is not always pronounced as /nra/, but also sometimes has the pronunciation /nta/. Thus, disambiguating the renderings ന്റ and ൻറ on the basis of pronunciation is quite unjustifiable.

The encoding for ന്റ is ന + ് + റ, following the Dravidian scheme for conjunct derivation. In the Dravidian scheme, there is an alveolar class comprising stop and nasal. Just as in Tamil, the alveolar stop uses the same glyph as റ. In Tamil, the alveolar nasal has a unique glyph. However, in Malayalam, the same glyph as dental-na is used for denoting alveolar-na.

In the case of ന്റ, it is a conjunct just as Dravidian ങ്ക, ഞ്ച, ണ്ട. ന്റ is a conjunct of alveolar-na and alveolar-stop.

Since the chillu-na is always pronounced as vowelless alveolar-na, the chillu glyph is an accurate representation in the ന്റ glyph. Thus, ന്റ is generated in a conjunct i.e., ന (alveolar-nasal) + ് + റ (alveolar-stop). This is expected by users since visually, ന്റ also looks much more like a conjunct than ൻറ.

On the other hand, ൻറ is chillu-ന followed by റ, but not as a conjuct, exactly like ൻസ, ൻക, etc. Therefore, the accurate encoding is na + chandrakkala + ZWJ + rra. This is expected by users since visually, ൻറ is chillu-ന followed by റ on a single line just as the other non-conjuncts involving ന.

If chillu-ന is encoded, it can be used either for ന്റ or ൻറ, but not both. There will be considerable confusion about which sequence uses the chillu-ന codepoint. This also causes immense problems for data entry operators and general users of Malayalam.

3 Recommendations

1. ന്റ = ന + ് + റ

2. ൹ = ന + ് + ZWJ + റ

4 Chillu Polyvalency

1 Background Information

Chillus are polyvalent and a single base consonant cannot be determined for a chillu. ർ can be derived from ര and റ, ൽ can be derived from ല, ത,and ദ and ൾ can be derived from ള, ട and ഴ. Since none of these chillus can be derived from a single base consonant, it is necessary to encode them as atomic codepoints.

2 Observations

The delegates observed that this argument is extremely hypothetical and does not demonstrate any actual problems in Unicode enabled applications. The existing encoding has several advantages to meet the needs of low-level Unicode applications as well as being sufficient for high-level applications.

The delegates noted that Malayalam has a given character set. This character set has evolved and has been used in a historically justifiable manner. In this character set, there are some characters which have multiple manifestations for a single character and a single manifestation for multiple characters. We have to encode the existing characters in a defined way by fixing each character at a specific place. The total etymology and grammar of each character does not affect the encoding of that character. On the other hand, each character must be able to achieve its natural behaviour in various contexts in computing applications.

Thus, it is necessary to define a single place for each such character. Conjuncts must have a single derivation from base characters; so should all presentation forms, including chillus, consonant signs, etc.

Chillus are requested to be encoded as atomic codepoints due to polyvalency, for e.g., ർ is the chillu of ര or റ, ൽ is the chillu of ല, ത and ദ, and ൾ is the chillu of ള,ട and ഴ. However, in Unicode applications, the

chillus cannot be placed in direct relationship with all their supposed bases. For e.g., in sorting, the chillu ർ cannot be placed before ര and റ, chillu ൾ cannot be placed before all of ല, ത and ദ, chillu ൾ cannot be placed before all of ള, ട and ഴ. Similiarly, for IDN, a chillu cannot be equal to all combinations of Consonant+chadrakkala, i.e., the chillu-ൾ cannot be equal to all of ള+്, ട+് and ഴ+്.

Both in linguistic theory and for proper operation of Unicode applications, it is necessary to fix the derivation of a chillu from a single base. The derviations, ർ from ര, ൽ from ല and ൾ from ള are not only linguistically justifiable, they also meet all requirements for Unicode applications including, fallback rendering, IDN and sorting. It should also be noted that the most used keyboard layouts follow the Inscript model of inputting a chillu as, base-consonant + chandrakkala + nukta. Hence, the selection of the single bases are well in line with user expectations and have great frequency of use.

3 Recommendations

1. Derive chillus from base characters as follows:
    1. ർ : ര
    2. ൽ : ല
    3. ൾ : ള
    4. ൻ : ന
    5. ൺ : ണ

5 Stripping of Joiners

1 Background Information

ZWJ and ZWNJ are so-called formatting control characters in Unicode. In general, they are used to effect changes to the default rendering of seqences of codepoints. In the Unicode standards, it is mentioned that these codepoints may be stripped in applications for certain kinds of processing. The contention is that stripping the joiners causes semantic changes in the text. Issue was taken also with the fact that the joiners are ignorable in Unicode.

2 Observations

The participants observed that applications are free to give value to the joiners, or ignore them totally, depending on the requirements of processing. The ignorability of joiners is sometimes required for the correct processing of text.

For e.g., in IDN, the ZWJ and ZWNJ are mapped to empty string which reduces the possibility of spoofing. This could be seen as ignoring ZWJ and ZWNJ.

Also, for e.g., in rendering programs like text editor, ZWJ and ZWNJ affect text rendering. This could be seen as not ignoring ZWJ and ZWNJ.

Completely deleting the ZWJ and ZWNJ from the pristine data is not expected in most applications, and such applications should be considered as not supporting Malayalam.

The delegates considered the problems of ZWJ/ZWNJ in collation. In sorting, the ZWJ/ZWNJ must be given the ignorable value, such that the sequences ന്മ and ന്‍മ have the same primary value. The delegates also noted that chandrakkala need not be treated as samvruthokaram unless there is a reliable way of determining it contextually.

The delegates also expressed their concern over the data given in L2/06-189 regarding chillu and chandrakkala which is absolutely misleading. The ambiguity between chillu and chandrakkala that emerged in

language is a product of the script reformation. The reason that newspapers and publishers do not use the samvruthokaram, is because of reformation.

In L2/06-189, several samples are given of contrastive examples of chillu and chandrakkala. However, none of these examples pose a problem in low-level or high-level applications.

The difference suggested in the document that കാല് means leg and കാൽ means one-fourth, is absolutely wrong; these two words are one and the same. Thus, there is no semantic distinction to be maintained between them.

As per present Unicode norms, to use the Original form of the samvruthokaram the sequence ു + ൂ may be used and no additional facility is required.

3 Recommendation

1. ZWJ/ZWNJ is by default ignorable in collation, and may be given value in combination with other codepoints
2. ു + ൂ may be used to represent the original form of samvruthokaram.

6 Character Repertoire

For Unicode encoding, it is very important to determine a character set comprising of vowels, consonants, signs and conjuncts, because:

1. The very basis of Unicode technology is to encode only basic characters and conjuncts are generated from these basic characters. Due to this, it is important to determine and fix the set of basic characters namely vowels and consonants.
2. It is also important to unambiguously determine the combination of basic characters which generates a conjunct. E.g., മ്പ = മ + ◌ൂ + പ and not ന + ◌ൂ + പ.
3. The visual form of a conjunct must also be determined സ്ന, സ്ഥ, ക്ത, ന്റ
4. The conjuncts formed by consonant signs must also be determined as in the case of other consonants, e.g. കൂ, ക്യ, ക്ല etc.
5. The identity or value of anusvaram, chandrakkala, praslesham, visargam, etc should also be determined and fixed.
6. The order of the above basic characters and glyphs (alphabetical order) should also be fixed.
7. It is a technical requirement that in order to maintain the intent of the author an optimal set of characters should be included in every font, due to fallback rendering.
8. There should be a specfic norms for conjunct forms, i.e., if two consonants form a conjunct, then it should be treated as a conjunct and its visual appearance should be fixed.
9. Those conjuncts which have been historically evolved and found to be in common use should be accepted.
10. A consonant forms conjunct with many consonants. Standardising some of those combinations, but rejecting others and splitting them with chandrakkala, even though all the combinations are widely used in the language, is not acceptable. For e.g., considering the conjuncts of സ, we cannot accept സ്ന, സ്ത, സ്ഥ, and at the same time, reject സ്ക, സ്പ, സ്മ. It is quite irregular and therefore unacceptable.
11. Consonants form conjuncts by joining in some definite ways:
    1. subjoining the second consonant under the first consonant
    2. joining the second consonant on the right of the first consonant
    3. combination with complete change of form

    We have to fix the different conjuncts in the above mechanisms the same way as they have naturally evolved in the script.

The "Report of the Committee on Malayalam Character Encoding and Keyboard Layout Standardisation" [1] was discussed, and several delegates were of the opinion that this report contained too many factual and procedural errors and that a fresh report should be submitted after due consideration of facts.

Some delegates pointed out that, even though the Unicode standards do not permit the deletion of characters, this report deletes several of them, such as some vowels, indigenous numerals, etc.

The delegates pointed out that this was repeated in the case of the submission of filled form for proposals for amendments to the ISO-10646, ignoring the recommendation on the form itself that codepoints may not be deleted. They also noted that, the report makes no mention of ZWJ/ZWNJ even though they are an integral part of the Indic encoding model.

The workshop resolved to appeal to the Government to submit a detailed document for the logical encoding of Malayalam in Unicode.

From a technical perspective, it is necessary to have a commonly agreed character set including conjuncts in order to maintain usability of Computer User Interfaces. This is particularly true in the event of fallback rendering.

Fallback rendering is a mechanism to produce legible and acceptable displays of text in all conditions. Under fallback rendering, for a given sequence, a priority order for acceptable display is chosen and used in the presentation layer of the software. Thus, the chillu form of ൻ is rendered for a sequence ന + ് + ക.

In the case of ദുക്സാക്ഷി, if a computer with a font F_1 which does not have the ക്സ conjunct, is used to type ദുക്സാക്ഷി, then with an encoding of ക + ് + സ the stylistically correct rendering is obtained. However, when the text is transmitted to computer with a font F_2, which does have the ക്സ conjunct then the same encoded text gives a stylistically incorrect text (ക്സ conjunct) although it is still quite legible.

In order to prevent such a situation, the only mechanism is that in computer with F_1, there should either be a provision for automatically adding ZWNJ to force the visible chandrakkala rendering, or F_1 itself should contain the ക്സ which acts as feedback to the user to type ക്സ . Therefore, an optimal set of such conjuncts

1 submitted to UTC and recorded as L2/07-013

should be available in fonts to achieve practical interchange of style information.

On the basis of the above norms, the workshop scrutinised the complete character repertoire, one by one and prepared a list of valid Malayalam glyphs.