L2/07-035

Atomic Chillus causes Spoofing

Rajeev J Sebastian
Rachana Aksharavedi
(Presented at the workshop “Problems of Malayalam encoding in Unicode”
held at Univ. of Kerala, January 24-25, 2007)

L2/06-189 claims that the use of ZWJ to represent chillus makes it impossible to register certain domain names. It further claims that the encoding the chillus would alleviate this problem. Further, claims have been brought that there is a contrastive use of chillus due to which unique domain labels would overlap preventing the use of one of the overlapped domains.

In the following sections, we analyse these claims showing that the use of ZWJ is certainly possible in domain names, and further that their use is highly advantageous to Malayalam. We also show that encoding chillus and/or giving a non-ignorable status to ZWJ/ZWNJ causes serious spoofing opportunities.

IDNA

In IDNA, a domain label is processed using Nameprep and Punycode to generate an domain name in ACE. Nameprep is applied to reduce variant forms, confusables, etc to a unique label. This generated label is then processed by Punycode to generate a final ACE form.

We needn't consider the Punycode algorithm since it merely mechanically transcodes the Nameprepped Unicode label into an ASCII Compatible Encoding. Thus, if the Nameprepped labels are the same, then their corresponding Punycode labels are also the same.

The Nameprep process consists of 4 steps:

1. Mapping
2. Normalization
3. Prohibit
4. Check-bidi

In the case of the chillu issue, only Mapping and Prohibit need be considered. The current chillu encoding is neither affected by Normalization, nor by Check-bidi.

In the Mapping step, each codepoint is checked against some tables in the Nameprep: relevant table is Table B.1 from Appendix B of Stringprep.

The output of the Mapping step is then normalized.

The normalized label is then checked for Prohibited characters specified in Nameprep: relevant table is Table C.2.2 from Appendix C of Stringprep. If a prohibited character is present in the domain label, then the process exits and returns an error, because prohibited characters are not permitted in the domain name.

Additional Mapping and Prohibit tables may be defined by registrars or other authorities for specific purposes such as for specific TLDs, etc.

IDN for Malayalam

In the case of IDN for Malayalam, there is no need for additional mapping tables. Since Table B.1 maps ZWJ and ZWNJ to the empty string, the 3 different manifestations of a C1 + chandrakkala + C2 sequence map to the same domain. Also, ZWJ never appears in the Prohibit step of Stringprep, where the joiners throws an error.

Rendering	ന്മ	ന്‍മ	ന്‌മ
Encoding	ന ◌് മ	ന ◌് ZWJ മ	ന ◌് ZWNJ മ
Mapping	ന ◌് മ	ന ◌് മ	ന ◌് മ
Normalization	ന ◌് മ	ന ◌് മ	ന ◌് മ
Prohibit	ന ◌് മ	ന ◌് മ	ന ◌് മ
Check Bidi	ന ◌് മ	ന ◌് മ	ന ◌് മ
Punycode	xn--uwcm6g	xn--uwcm6g	xn--uwcm6g

Table 1: 3 equivalent manifestations generate equivalent Punycode in the current system

L2/06-189

L2/06-189 argues that in since ZWJ and ZWNJ are Prohibited, domains such as സര്‍ക്കാര്‍ are not registerable. This is obviously a false statement, since the sequence സര്‍ക്കാര്‍ will be mapped to the ACE label xn—bwca3fc0b1bygbd, which establishes that it is registrable.

```
Python 2.4.4 (#1, Oct 25 2006, 16:27:12)
[GCC 3.4.6] on linux2
Type "help", "copyright", "credits" or "license" for more information.
>>> u'സര്‍ക്കാര്‍'.encode('idna')
'xn--bwca3fc0b1bygbd'
>>>
```

The reason for this statement in L2/06-189 is the attached note from a Registration Authority, that ZWJ/ZWNJ are prohibited from the DNS. Apparently, this was in response to some submitted mapping tables (which the authors do not reveal). The authority also asked the authors to review the relevant RFCs before submitting proposals for mapping tables.

The authors interpreted this admonition from the authority, and consequently the Table C.2.2 of Stringprep, as a statement against the use of ZWJ/ZWNJ in domain names. It is easily seen that the authors were quite wrong in their interpretation. Rather, it shows that they have not at all examined IDN, because it can be seen that the use of ZWJ in domain names is not only possible, but in fact advantageous for Malayalam.

IDN should be considered from the totality of its implementation. The IDN environment contains a client (such as a web browser), and a DNS server. It is important to note that the DNS servers only handle the ACE domain labels and not Unicode strings.

The Unicode strings are rendered at the client using the Unicode shaping engine. Thus using ZWJ and ZWNJ in the URL text widget of a browser allows correct rendering of a domain name. The entered domain name is processed using Nameprep and Punycode, and it is the result of this processing that is resolved by the DNS resolver using data from the DNS servers. Hence, the ZWJ/ZWNJ may be used in domain names.

This particular argument demonstrates the total lack of understanding of the expected implementation of IDN, and also shows that the authors do not care to check their facts, or relevant standards.

വൻയവനിക vs വന്യവനിക

Another argument raised for the chillu encoding was the emphatic claim that വൻയവനിക and വന്യവനിക should map to 2 different domain names, since there is a semantic difference between those 2 sequences.

It is important to note that neither of these contrived examples are accompanied by proof of their existence or usage in Malayalam. In fact, these sequences do not meet the rules of Malayalam word-formation. Therefore considering them by the “meaning” of these words is quite unacceptable.

However, it is useful and possible to consider them merely as sequences which lead to differing renderings without considering any proposed meanings: ൻയ in one case, and ന്യ in another.

It is important to note that in Malayalam the sequences ന്മ, ൻമ and ന്‌മ are equal in value. They are merely manifestations of the same underlying cluster ന + ◌് + മ. In Malayalam, although there are three manifestations of a cluster, in general, words with different meanings do not occur in which the only point of difference is the particular manifestation of the same cluster.

Also, in Malayalam, the rendering ന്യ is considered as a conjunct similiar to ന്മ. So, ന്യ and ൻയ should also be considered equal in value, just as in the case of ന്മ and ൻമ. This can be quite easily seen in the case of sorting, where both are sorted at exactly the same place, with exactly the same value. In the case of IDN, just like sorting, these sequences should map to the same Punycode.

The reality is that Malayalam, just like most other Indic scripts, is a Complex Text Language (CTL) in which a codepoint or sequence of codepoints may create different renderings based on their position in the text stream, or even on different computers equipped with different fonts. There is no such thing as a “rendering equivalence”.

Also, as sequences, it is important to note that it is not at all unreasonable for words with different meanings and Unicode encodings to generate exactly the same domain name. In the IDN environment, where words may be chained to generate identifiers, it is not possible in any language to avoid such occurrences. Several such examples have been shown[1]. In such situations of language use, it is necessary to use other methods to select appropriate non-conflicting domain names, such as using hyphens to seperate such words.

In the case of Malayalam, it is quite rare if not impossible for such examples as വൻയവനിക and വന്യവനിക to occur naturally. However, as part of the language mechanism, if the person wishing to register വൻയവനിക.com finds that it is unavailable, he would select വൻ-യവനിക.com instead. In fact, in this particular case, it is much more appropriate to use a hyphen than not. Of course, the persons suggesting വൻയവനിക chose to portray it as a word, rather than a badly formed expression, in order to mislead people in other areas of the chillu encoding debate.

[1] Examples include, powergenitalia.com (Powergen Italia, an Italian power-related company), penisland.com (Pen Island, a pen manufacturing unit), expertsexchange.com (a website dealing with computer experts, advice, support and consulting), etc. Also see, http://www.snopes.com/business/names/powergen.asp

Spoofing using the chillu encoding

The definitely much larger number of cases in Malayalam words where the different manifestations do not affect the meaning at all leads to a very serious problem if the chillus are encoded, or if the joiners are given some non-ignorable status in IDN as in PRI-96.

Due to backwards compatibility, the sequence ന + ◌് + ZWJ + മ must continue to be rendered as ൻമ, irrespective of the chillu encoding.

This causes a major security issue, because the mixed use of chillu codepoints and ZWJ-based chillus can open a huge opportunity for spoofing and phishing attacks.

In the table below, we can see that the punycodes generated for the sequences are quite different:

Rendering	ന്മ	ന്‍മ	ന്‌മ	ൻമ
Encoding	ന ◌് മ	ന ◌്ZWJ മ	ന ◌്ZWNJ മ	ൻ മ
Mapping	ന ◌് മ	ന ◌് മ	ന ◌് മ	ൻ മ
Prohibit	ന ◌് മ	ന ◌് മ	ന ◌് മ	ൻ മ

Table 2: First three columns vs the last column illustrates spoofing issues with atomic chillus

Thus words with the exact same meaning, നന്‍മ and നന്മ may cause to have different Punycode encodings, leading to spoofing attacks. If a word, or sequence, contains 'n' chillus, then it is possible to generate atleast 2^n domains which do not differ in rendering, but does differ in encoding. It is also important to note that there are a very large set of possible sequences and combinations of consonants which do not have conjunct forms, and this can be exploited in conjunction with Fallback Rendering.

സര്‍ക്കാര്‍	സർക്കാർ	സര്‍ക്കാർ	സർക്കാര്‍
സ ര ◌്ZWJ ക ◌് ക ◌ാ ര ◌്ZWJ	സ ർ ക ◌് ക ◌ാ ർ	സ ര ◌്ZWJ ക ◌് ക ◌ാ ർ	സ ർ ക ◌് ക ◌ാ ര ◌്ZWJ

Table 3: 2^n possible domains with aotmic chillus when considering backwards-compatibility of shaping engines

It should be noted that in English too, spoofing attacks may occur in certain cases. The Paypal spoofing example is famous and memorable, because using the upper-case L and the lower-case I looks very similar with certain fonts: paypaI and paypal.

However, in Malayalam, any font which contains a chillu (which is required in all Malayalam fonts), leads to the exact same rendering in any of the combinations in the above example of സർക്കാർ. Thus, spoofing requires only correct implementations of Unicode and proper text shaping.

Case for ZWJ/ZWNJ in IDN

The reason for this is the equivalence of the 3 manifestations of the C_1 + chandrakkala + [ZWJ/ZWNJ] + C_2 sequence as mentioned earlier. In all cases, words using these sequences are entirely equivalent. It is rare, if not impossible, for a counter-case with a possible word in Malayalam language. This also reflects on any move to give a non-ignorable status to ZWJ/ZWNJ in Malayalam such as in PRI-96. We strongly advise the UTC not to accept the PRI-96, atleast in the case of Malayalam.

Rendering	യൂണീക്കോഡ്	യൂണീക്കോഡ്
Encoding	യ ◌ൂ ണ ◌ീ ക ◌് ക ◌ോ ഡ ◌്	യ ◌ൂ ണ ◌ീ ക ◌് ക ◌ോ ഡ ◌്ZWNJ
Mapping	യ ◌ൂ ണ ◌ീ ക ◌് ക ◌ോ ഡ ◌്	യ ◌ൂ ണ ◌ീ ക ◌് ക ◌ോ ഡ ◌്ZWNJ
Prohibit	യ ◌ൂ ണ ◌ീ ക ◌് ക ◌ോ ഡ ◌്	യ ◌ൂ ണ ◌ീ ക ◌് ക ◌ോ ഡ ◌്ZWNJ

Table 4: The word 'Unicode' in Malayalam

Rendering	സ്‌റ	സ്‌റ
Encoding	സ ◌് റ	സ ◌്ZWNJ റ
Mapping	സ ◌് റ	സ ◌്ZWNJ റ
Prohibit	സ ◌് റ	സ ◌്ZWNJ റ

Table 5: Hypothetical cluster സ്‌റ to illustrate spoofing issues in PRI-96, independent of chillus

Rendering	ന്യ	ന്യ
Encoding	ന ◌് യ	ന ZWJ ◌് യ
Mapping	ന ◌് യ	ന ZWJ ◌് യ
Prohibit	ന ◌് യ	ന ZWJ ◌് യ

Table 6: Spoofing issues in PRI-96, in conjunction with PR-37

Usually, it is appropriate to select one particular rendering for purposes of style. However, it is not mandatory, and the selection of appropriate manifestation is dependent on culture, knowledge of the user, etc. There is sufficient evidence to show that earlier in the history of the language, it was more appropriate to write ദൃക്സാക്ഷി than the modern style of ദൃക്‌സാക്ഷി.

Hence, glyph selection is dependent on cultural factors than on questions of meaning, because meaning does not even enter into the picture. Of course, persons with some vested interests may always try to bring in such contrived examples; it is possible to do so in every script, not just in Malayalam.

The equivalence of value of the three manifestations is quite apt as a general rule for low-level applications of Unicode. This is especially true in the aftermath of the disastrous script reforms of Malayalam.

As a result of the script reforms, whatever existing consistent scheme of conjunct formation and use was lost, and renderings involving chillus and overt-chandrakkala became equivalent to conjuncts composed of the same basic characters. Software and even Govt standards mandated different sets of conjuncts haphazardly with no linguistic basis for their selection.

Irrespective of the scheme of conjunct formation before or after the reform, at the encoding level it is not at all necessary to give atomic encodings to C1-conjoining forms (chillus and the special case of ക്‍), C2-conjoining forms (consonant signs), or conjuncts. The user selects the appropriate rendering using the ZWJ/ZWNJ, which reflect exactly that the user wishes to use a different rendering which has equal value to the default rendering, but may be better in that particular situation, or may reflect the concious or unconcious choice of the user.

This is particularly true in the case of transmitting URLs, say over a telephone. Since it is equivalent to write നന്മ.com and നന്‍മ.com, it is entirely possible that users will type the wrong domain and thus visit an unintended website.

Chillus and Polyvalency

The polyvalency argument for encoding chillus is quite inappropriate from the point of extant linguistic studies of Malayalam. Please see statements from Malayalam language experts Chitrajakumar and Gangadharan, and others on this. It is because of the polyvalency argument that the tentative atomic chillus cannot be mapped to a single vowelless base consonant.

Furthermore, from the point of view of input methods, a much larger proportion of current users of Malayalam software products are aware of and use chillus as if they are derived from the single base consonants. In the most popular product ISM Gist developed by CDAC, and in most other packages implementing the Govt mandated Inscript keyboard scheme, the sequence of inputting a chillu is quite similar to the ZWJ encoding, i.e., as consonant + chandrakkala + NUK.

From the point of view of other low-level applications, the chillu encoding cannot solve questions relating to its underlying value, such as their positions in the sort order and their use in IDN. The underlying model of single-base chillus is well-adapted to such questions. Moreover, the particular choices for single-base, i.e., ര for the ര്‍ chillu, ല for the ല്‍ chillu and ള for the ള്‍ chillu provide elegant solutions for these questions, while not impacting higher-level applications in any way at all. In conjunction with the 3-point solution proposed by Rachana, the Malayalam encoding becomes computationally efficient, linguistically sound, harmonious with other Indic scripts and the virama model, usable and matches with legacy applications.

Thus, it is entirely appropriate to use the ZWJ encoding for chillus. These applications demonstrate the underlying value of sequences and different renderings.

Atomic chillus alone do not solve the problems cited by the proposers. It would have to be accompanied with the removal of ZWNJ from Malayalam, removal of PR-37 and removal of many levels in Fallback rendering; this is tantamount to abandoning the virama model, and moving to a pure visual model for encoding Malayalam which is quite inefficient, creates confusion, prevents backwards-compatibility (and hence the Stability Policy), and finally makes it extremely unusable for users of the encoding.

In conclusion, the atomic chillus do not really solve any problems of Malayalam, and increase confusion and introduces new security issues. This severly impairs the use of Malayalam in computing environments.