Notes on PRI-96 feedback
Rajeev J Sebastian

Following I present some notes on the feedback provided by Cibu on his webpage http://unicode.wikia.com/wiki/Malayalam/PR_96:_feedback

**Case 1:**

Column 2 ആണതരി, Column 4 ഉലഘാടനം, and Column 5 ആളകരം, are wrong. This is because due to fallback rendering, removal of ZWJ will lead to the exact same rendering: there is no conjunct for ണ്‌ത + ത, for ണ്‌ത + ത and nor for ല + ഘ+ ല, for ല + ഘ for ചഞ്ചം, for ചഞ്ചം + ക+ ക.

Hence, ആണതരി will fallback to ആണതരി and not ആണ്ടരി. Similarly in other cases.

[On a side note, no word ആണതരി is possible, since ണ്‌ത + ത, will become ണ്‌ത + ത + ട (ക്കാലി, + ട = ക്കടാ) in sandhi. This is a regular morpho-phonemic rule of Malayalam. ]

**Case 2:**

The examples given in case 2 are all “wrong”. They should all be written as 2 words:

േേശ്‌ രാഗം
ഓണ്‌ പായ്
ഓണ്‌ യനം
ൊൊൊൊounces (House Warming)
ഓൊൊ (Arab Bazaar)
േപ്പിംഗ്‌ കണം (Paying Counter)

Cibu also noted that “in Malayalam, space is not mandatory between a noun and its adjective noun”, and gives an example മലയാളവ്ബ്കരണം which is equivalent to മലയാളവ്ബ്കരണം.

However, unlike this example, the above mentioned examples are different: they are not all Malayalam words and as such, they follow different rules when used in sentences and when considering joining.

In fact, the scan provided by Cibu is quite appropriate here. It can be roughly translated as “Thakazhi and Bashir know Malayalam grammar. The only thing is, they have not spent any effort to learn about Malayalam grammar. Knowing grammar and knowing about grammar are two different things.”. Being a Malayalee, Cibu knows some grammar. However, from this we cannot extrapolate and consider that he knows the underlying theory of grammar, or the implications of grammar from all perspectives, or of language use of Malayalees. This is what professional linguists study.
Case 3 and Case 4:

Historically, ദകാകി was an acceptable rendering for ദക്‌സാകി. In modern times, it is more stylish to write it as ദക്‌സാകി. In some circumstances or styles, some printers may choose to use ദകാകി. It shows that conjunct formation is highly dependent on culture and style of the writer. Both cases 3 and 4 show explicitly the problems of giving ZWJ/ZWNJ a mapping different from null string (or empty string). Two sequences different in encoding and in Punycode but with same rendering (perhaps dependent on style) and meaning, will lead to spoofing issues.

Finally, it is important to note that Malayalam should not be played with for the sake of argumentation. It is meant for use by large number of people. Cibu is advised to learn Malayalam and understand the problems of Malayalam use before commenting on the same.

When his samples വനയവനിക and മനവിേകാഭം were presented at the workshop at Univ. of Kerala by the organizers, it evoked much laughter and many comments among the learned participants. I will not reveal the comments made by the delegates, so that the sentiments of the creator of these “words” is not hurt. The samples given above are of the same nature.

Also, Cibu talks about removal of ZWJ/ZWNJ. I do not understand how this is a problem. Modern Internet transmission formats are well-designed and do not require deletion of ZWJ/ZWNJ in transit. It should also be noted that IDN's can always be transported in a UTF, and transformed to Punycode only just before DNS resolution. The ability of Internet transports to carry UTF is not questioned; for e.g. both HTML and MIME can do it. Of course, for processing, it may be necessary to disregard some codepoints.

When analysing domain names, we do not analyse them from the point of view of words, rather we analyse them from the point of view of sequences. E.g., cnn in cnn.com is not a word. However, we do know that CNN.com is exactly the same as cnn.com. The easily visible distinction between CNN.com and cnn.com affects other examples. However, it is a function of the protocol for use of IDNs.

Irrespective of chillu encoding, (or expected encoding of conjuncts and C2-conjoining forms), the underlying linguistic sequence in വനയവനിക (ക്കക) is the same as വന്വനിക (ക്ക). Average users recognize this principle and will use it in a language interface. Thus, any mapping of ZWJ/ZWNJ or any encoding of conjuncts, C1-conjoining forms and C2-conjoining forms will confuse users and lead to security issues.

It is also against the principles of computing to delete pristine data. Applications that randomly delete data should not be considered as supporting the transmission of Unicode data (and in general should be considered as buggy applications).

Also, standards cannot built on a house of cards: if standards are to assume faulty implementations, then the standard has been terribly unsuccessful. However, standards can greatly simplify implementations via the choices made by them; atomic chillus unfortunately greatly increase complications in applications.

---

1 The distinction between ExpertsExchange.com and ExpertSexChange.com is lost upon case folding. However, Punycode provides a method for transporting these distinctions in the Punycode encoding without affecting DNS resolution, i.e., ExpertsExchange.com can be transported without losing case information, but at the same time, it is folded with other similar sequences. It would be good for UTC to investigate methods of achieving this for Indic IDNs using this existing scheme.