L2/07-061

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:

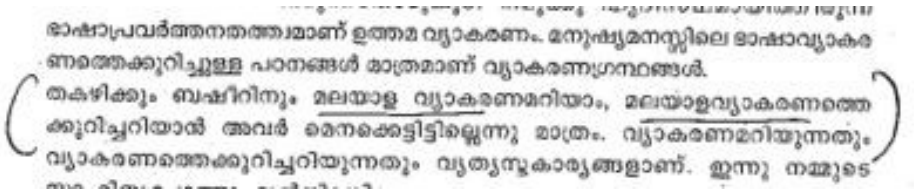
Column 2 ആൺതരി, Column 4 ഉൽഘാടനം, and Column 5 ആൾക്കൂരങ്ങ്, are wrong. This is because due to fallback rendering, removal of ZWJ will lead to the exact same rendering: there is no conjunct for ണ + ് + ത, for ല + ് + ഘ nor for ള + ് + ക + ് + ക.

Hence, ആൺതരി will fallback to ആൺതരി and not ആണ്തരി. Simliarly in other cases.

[On a side note, no word ആൺതരി is possible, since ണ + ് + ത, will become ണ + ് + ട (ണ്ട i.e., ആൺ + തരി = ആണ്ടരി) 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:
ദേശ് രാഗം
പ്രാഗ് യന്ത്രം
ഹൌസ് വാമിംഗ് (House Warming)
പാക് കരാർ
അറബ് ബസാർ (Arab Bazaar)
പേയിങ് കൌണ്ടർ (Paying Counter)


ഭാഷാപ്രവർത്തനതത്പരമാണ് ഉത്തമ വ്യാകരണം. മനുഷ്യമനസ്സിലെ ഭാഷാവ്യാകര
ണത്തെക്കുറിച്ചുള്ള പഠനങ്ങൾ മാത്രമാണ് വ്യാകരണഗ്രന്ഥങ്ങൾ.
തകഴിക്കും ബഷീറിനും മലയാള വ്യാകരണമറിയാം, മലയാളവ്യാകരണത്തെ
ക്കുറിച്ചറിയാൻ അവർ മെനക്കെട്ടിട്ടില്ലെന്നു മാത്രം. വ്യാകരണമറിയുന്നതും
വ്യാകരണത്തെക്കുറിച്ചറിയുന്നതും വ്യത്യസ്തകാര്യങ്ങളാണ്. ഇന്നു നമ്മുടെ

Cibu also noted that "in Malayalam, space is not mandatory between a noun and its adjective noun", and gives an example മലയാള വ്യാകരണം which is equivalent to മലയാളവ്യകരണം.

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, ദൃക്സാക്ഷി was an acceptable rendering for ദൃക്സാക്ഷി. In modern times, it is more stylish to write it as ദൃക്സാക്ഷി. In some circumstances or styles, some printers may choose to use ദൃക്സാക്ഷി. It shows that conjunct formation is highly dependent on culture and style of the writer. Both cases 3 and 4 shows 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 വൻയവനിക and മൻവിക്ഷോഭം 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 analysig 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[1]. 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 വൻയവനിക (നൻമ) is the same as വന്യവനിക (നന്മ). 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.