L2/13-036

Title: Old and New Chillus in Malayalam and implications for Sinhala

Author: Roozbeh Pournader and Cibu Johny (Google)

Date: 2013-01-29

Action: For UTC's information

Background

In UTC #133, we were asked to research the uptake of chillu characters in Malayalam.

We looked at Google's web corpus, comparing the number of web pages where common Malayalam words using a chillu form were encoded using pre-5.1 sequences of <consonant, virama, zwj> to the web pages where they were encoded using post-5.1 atomic chillu characters.

Here's a comparison of the frequencies of each encoding for the top ten or so most common words containing chillu forms (all counts have been multiplied by an arbitrary factor):

word	pre-5.1 encoding	5.1+ encoding	either form (union)	both forms (intersection)	uptake ratio of 5.1+ encoding
ആർ	30,742	4,670	34,776	636	12%
കൂടുതൽ	21,074	6,629	27,483	221	24%
മുതൽ	21,736	2,435	23,890	281	9%
ഇന്ത്യൻ	16,282	2,540	18,338	484	12%
എന്നാൽ	14,813	1,924	16,503	234	10%
ഞാൻ	14,104	3,085	16,483	706	15%
ഇപ്പോൾ	10,712	1,167	11,719	159	9%
ജൂൺ	2,589	6,776	9,355	10	**72%**
പേരിൽ	7,782	1,287	8,971	98	13%
അതിൽ	5,571	943	6,394	119	13%
അവൻ	3,127	448	3,527	48	12%

In calculating the uptake ratio, pages were both forms have been used have been ignored. The median for the uptake of the new chillu character among these most frequent words is 12%.

An interesting outlier where the new encoding is more popular is the Malayalam word for the month of June, ജൂൺ, most probably led by usage in software generating dates automatically based on CLDR data, and the frequency of dates in the Malayalam Wikipedia and its copies, both of which have adopted the new encoding.

It's noteworthy that some of the pages where the sequences appeared were Wikipedia articles in languages other than Malayalam. In those, like various other places on the web, the usage appears to be mixed. For example, as of the 2012-12-26, the English Wikipedia's article on GMail contained a tooltip using <la, virama, zwj>, while the article on Yahoo! Mail used the new chillu character U+0D7D <chillu-l> for the spelling the same chillu form.

Other observations

One of the authors have been observing the Malayalam web in Unicode since its early days, including the encoding of the atomic chillu characters in Unicode 5.1 and its aftermath.

In hindsight, we are not sure if we would do similarly and request the encoding of chillu forms if we knew what would happen in the Malayalam computing community.

For example, there is a sizable community of influential Malayalam software/font developers and content creators that actively resist supporting the new encoding of chillu characters. The reasoning for some of these preferences is partially based on ideological differences in the analysis of the Malayalam language and its orthography.

Also, the uptake numbers for the new encodings is potentially mostly driven by its adoption in the Malayalam Wikipedia and CLDR.

As time goes by, we expect the uptake to increase, but in practice, software implementations are required to support both encodings of Malayalam chillus for eternity, while no official equivalency is defined in the Unicode Standard. For example, rendering engines and fonts need to support displaying both forms, user-friendly text editors need to be able to treat the sequences encoded each way similarly, searching and collation engines are supposed to treat them the same way, etc. This gets more complicated as various pieces of software have bugs in handling zero-width characters, resulting in even more headaches for end users.

The situation is even less helped by the requirement in the Unicode Standard that software implementations may not silently convert the text they handle to text that is not canonically equivalent to the original. This means that copying and pasting text will guarantee the continuing existence of both encoding in Malayalam documents for the foreseeable future.

Conclusion

While the headaches for Malayalam is going to stay with us for a long time, we hope that the same would not happen for the Sinhala-speaking community. While encoding new pre-composed Sinhala forms may temporarily solve some display problems in some software, the long-term problems of software needing to support both encodings anyway is even more expensive. We believe that stabilizing the encoding of Sinhala as is serves its user community much better.

In the meanwhile, the authors will continue to push for better support for zero-width characters in software, especially in Google products. These characters are critical in representing the languages of hundreds of millions of users, from Persian, Urdu, Pashto, and Kurdish to Malayalam and Sinhala. They are also very important in preserving the exact meaning and presentation of text in several major scripts, ranging from Arabic, Devanagari, and Bengali to Khmer, Mongolian, and Buginese, and even non-complex scripts like Hebrew and Latin.