Additional IP Feedback related to Bengali Nukta

DATE 2018-08-28

# Overview

In testing the draft Bengali LGR against a set of putative labels, the IP has discovered an issue and would like to file this supplement to its feedback document “Response of IP to Bengali LGR Draft of August 9th 2018”, Dated 2018-08-14.

The IP would appreciate if the GP could review the findings below and make the appropriate updates to the document and the XML file when preparing the next version of the Bengali LGR.

# Concerns related to Bengali Nukta

The Bengali LGR restricts Nukta 09BC to a few consonants. These are
{0995-0997 099C-099D 09AB}

The IP has run the most recent (20180809) draft of the LGR over an existing corpus of about 77K Bangla labels with some interesting results. (The same test was repeated for 2K Assamese labels with similar results).

The context rule for Nukta is matched (corpus conforms to LGR) only for one context, that is the case  where 099C is followed by Nukta:

  Contexts matching rule "follows-only-specific-Cs":
    [099C] ⚓=[09BC]

However, several context were evaluated that lead to failing labels:

  Contexts not matching rule "follows-only-specific-Cs":
    [09A1] ⚓=[09BC]  - \*
    [09A2] ⚓=[09BC] - \*
    [09AC] ⚓=[09BC]
    [09AF] ⚓=[09BC]  - \*
    [09BF] ⚓=[09BC]

With exception of 09BF these are consonants that appear more frequently with NUKTA than the set formally allowed in the LGR.

The ones annotated with a \* are called "post-reform letters" in <https://en.wikipedia.org/wiki/Bengali_alphabet>.

Looking over the Unicode charts, we see that the three combinations are canonical decompositions for three characters that Unicode added under the heading "additional consonants" without further explanation.

According to the corpus I looked at, out of 77K labels, 56 had a wrong context for 09BC, the majority being [09AC] which is in fact not one of the three "reform letter" combinations.

Out of 77K labels the following where flagged with invalid sequences:

8285 had the sequence 09AF 09BC   (almost 11% of labels)
1793 had the sequence 09A1 09BC    (almost  3% of labels)
56 had the sequence 09AC 09BC
13 had the sequence 09A2 09BC
1 had the sequence 09BF 09BC

and only 67 labels contained a valid Nukta sequence 099C 09BC

That makes it appear that the rules for Bengali Nukta do not match actual usage.

The LGR proposal contains this statement, that we did not comment on, even though it has an issue:

"As there are provisions made in the UNICODE character table for Bengali, it is strongly recommended that  য় YYA (U+09DF), ড় RRA (U+09DC) and ঢ় RRHA (U+09DD) are to be used in Bengali LGR instead of “য”YA (U+09AF)+”়”Nukta (U+09BC), "ড"DDA (U+09A1)+”়” Nukta (U+09BC) and "ঢ"DDHA (U+09A2)+”়” Nukta (U+09BC). "

The "provisions" made by the Unicode Standard is to included several precomposed characters for "additional consonants". However, these consonants have a canonical decomposition and are also listed in the "composition exclusion" list - as a result, the decomposed(!) form is the one that is in NFC.

Root Zone LGRs (as well as IDNA2008) are presented in NFC, that is, 09DF , 09DC and 09DD simply ***cannot be used*** in the RZ-LGR. Instead, the sequences with Nukta ***must*** be used instead.

Whether 09AC and 09BF should also be allowed with Nukta is another issue. With slightly less than 0.1%, the case for 09AF is perhaps not the strongest and 09BF appears to be a single exception or even a typo. This is something that the language experts would need to decide.

***Conclusion:***

(1) the statement (cited above)  given in Section 3.3.6 of the document must be corrected to take into account the use of NFC for IDNA2008. (Note that user agents may allow un-normalized input by users, but that's outside the scope of the RZ-LGR).

(2) the list of special "C" to be used with Nukta needs to be reviewed. At the minimum, the three reform letter sequences must be added.

(3) the existing comments by the IP regarding Nukta and WLEs for Nukta remain unchanged and valid, and the IP expects those to be addressed as well.