IP Review of Bengali (Bangla) LGR Draft

DATE: February 18, 2019

VERSION reviewed: version 4.8.1 (xml and supporting document, except label file is 19oct15)

# Overview

The proposal appears to be in good shape and the IP sees no obstacle to taking this proposal to public comment. The IP tested against a corpus of 70K putative labels and found the rejected percentage very low (commensurate with noise/spelling error expected in the corpus). A couple of small issues are listed below.

# Comments

(1) There was no updated test label file, so we tested the 19oct15 version. That version has one label that is not normalized to NFC.

**Recommendation:** please normalize or pick a different test label.

(2) There are several small issues with the wording of the 2nd paragraph in section 3.3.6 Nukta that should probably be addressed by a slight rewording.

(2.1) The text uses "The UNICODE" which is not the preferred way of citing "The Unicode Standard".

(2.2) It is the definition of "composition exclusion" in Unicode Normalization Form C, not the character code tables, that define the sequences as the normalized form.

(2.3) The requirement that IDNs must use NFC is laid down in RFC 5891, not the Unicode Standard, and RFC 7940 mirrors this requirement for the specification of label generation rules.

We bring to the GP's attention the following suggestion for alternative wording:

The IDNA Protocol (RFC 5891) states that IDNs must be in Unicode Normalization Form C (NFC). RFC 7940 applies this requirement to LGRs. The definition of NFC in the Unicode Standard contains a number of composition exclusions. As a result, the Bangla letters য় YYA, ড় RRA and ঢ় RRHA have to be represented in the this LGR by using the sequences (YA +Nukta: U+9AF + U+09BC), (DDA + Nukta: U+9A1 + U+09BC), and  (DDHA + Nukta: U+9A2 + U+09BC) instead of the single code points YYA (U+9DF), RRA (U+09DC), and RRHA (U+09DD), although the  use of ‘Nukta’ is otherwise completely unnatural in Bangla.

**Recommendation:** suggest revising the wording, if possible