Bengali Script: Formation of the Reph and use of the ZERO WIDTH JOINER and ZERO WIDTH NON-JOINER

Written by: Paul Nelson, Microsoft Corporation

Overview:

In the process of implementing support for the Bengali script I have come across issues where ambiguity has been found. This paper is to provide suggestions for resolving the ambiguities and provide a starting point for public discussion to arrive at a standardized and normative description of how Bengali script should be implemented. The definitive use of the ZWJ and ZWNJ with Bengali script is critical for the stability of the implementation of shaping engines, lexical tools and ultimately the ability to exchange documents.

Reph and Yaphala

Reph: The formation of the Reph form is defined in the Unicode Book, Section 9.1, Rules for Rendering, R2. Basically, the Reph is formed when a Ra which has the inherent vowel killed by the virama/halant begins a syllable. This is shown in the following example.

র + ্ + ম ⟶ র্ম as in কর্ম

Yaphala: The Yaphala is a post-base form of Ya and I formed when the Ya is the final consonant of a syllable cluster. In this case, the previous consonant retains is base shape and the virama/halant is combined with the following Ya. This is shown in the following example.

ক + ্ + য ⟶ ক্য as in বাক্য

Issue: An ambiguous situation is encountered when the combination of Ra + virama/halant + Ya is encountered.

র + ্ + য ⟶ র্য or র‍্য

Proposed normative behavior: To resolve the ambiguity with this combination and to have consistent behavior, we need to look at the processing order of the Bengali script. When parsing the text, the ability to form the Reph is identified first and therefore the Reph form should have priority in processing. Thus, it is necessary to insert a ZWNJ character into the stream between the Ra and virama/halant to allow the virama/halant and Ya to be grouped together during processing. Thus, a normative solution to this ambiguous situation is proposed as follows.

র + ্ + য ⟶ র্য

র + ZWNJ + ্ + য ⟶ র‍্য

In the previous example, the ZWNJ is used because we are saying that we want two characters that would normally join to remain as separate entities.

Note: While in this situation a ZWJ may render the same results if rules for joining the Ra + ZWJ + virama/halant into a unit, it is important for a defined behavior of using the ZWNJ be used for this to communicate the correct meaning and to have uniform Unicode streams for this situation.

Other combinations with Ra and Ya

র + ্ + ZWJ + য ⟶ র্য (half form of ra)

র + ্ + ZWNJ + য ⟶ র‌্য (halant form of ra)