

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

Written by: Paul Nelson, Microsoft Corporation, 6/16/2003

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 (karma)

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 its 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 (bakyô)

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 ZWJ 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.

Formation of Bengali Ligatures

The Bengali script has special forms of consonants or syllables that combine with vowels to create ligatures. This ligation should be treated as conjuncts. By default, the vowel does not join with the preceding syllable. To enable this behavior, the ZWJ is used to force the joining behavior to happen in a situation where this would not normally occur. Thus, we have the following examples.

Normal Behavior

ক+্+র+উ=করউ

গ+উ=গউ

Ligated Form

ক+্+র+ZWJ+উ=ক্র

গ+ZWJ+উ=গু

Formation of Alternate Shapes

When typing Bengali script text, there are some consonant vowel combinations that take a special form by default. For example, Sha + Ukaar form a special form.:

শ + ু = শু

In the event a user would want the Ukaar to display below the Sha in other than the special form, the ZWNJ should be used to stop the joining behavior that happens by default. Thus, we have an example:

শ + ZWNJ + ু = শূ

Stand Alone Forms

When printing books, like dictionaries or for studying the language, there are occasions where combining marks or other forms are desired to display in stand alone form without showing a dotted circle. To achieve this display, the rendering engine needs to have a queue that the following sequence is valid. To achieve this, the use of the ZWJ as the first character in the “word” should be defined as the trigger. For example, a user wants to type a stand alone Yaphala. Simply typing the virama/halant + Ya will result with the halant sitting on a dotted circle and the Ya displayed as the first consonant of the next syllable cluster. By using ZWJ + virama/halant + Ya the rendering engine has a queue that beginning the sequence with the combining mark is valid and therefore the Yaphala may be rendered as desired.

্ + য = ্য

ZWJ + ্ + য = য়

The use of the ZWJ should be available for use with any characters that normally combine and may need to be displayed as an individual piece. While it is important for book makers to have this capability, the usage of this should be discouraged in normal running text as it has a destabilizing effect on lexical tools and impacts document interchange.