

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

Written by: Paul Nelson, Microsoft Corporation
Last Updated: 30 June 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.

U+09B0 U+09CD U+09AE

র + ্ + ঞ → ঞ 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 its base shape and the virama/halant is combined with the following Ya. This is shown in the following example.

U+0995 U+09CD U+09AF

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.

U+09B0 U+09CD U+09AF or U+09F0 U+09CD U+09AF

র + ্ + য → র্য or র + ্ + য → র্য

U+09B0 U+200C U+09CD U+09AF or U+09F0 U+200C U+09CD U+09AF

র + ZWNJ + ্ + য → র্য or র + ZWNJ + ্ + য → র্য

A ZWJ may render a ligated form of the consonant yaphala combination if a ligature form is found in the font. If the ligature form is not present in the font, it would render the normal yaphala form.

U+09B0 U+200D U+09CD U+09AF

র + ZWJ + ্ + য → র্য

Other combinations with Ra and Ya

U+09B0 U+09CD U+200D U+09AF or U+09F0 U+09CD U+200D U+09AF

র + ্ + ZWJ + য → র্য় (half form of ra)

ৰ + ্ + ZWJ + য → র্য় (half form of ra)

U+09B0 U+09CD U+200C U+09AF or U+09F0 U+09CD U+200D U+09AF

র + ্ + ZWNJ + য → র্য় (halant form of ra)

ৰ + ্ + ZWNJ + য → র্য় (halant form of ra)

Yaphala Behavior

As illustrated above, the yaphala is formed by typing U+09CD U+09AF. In the case of some letters, a ligated yaphala form may be desired. This is achieved by typing the ZERO WIDTH JOINER characters, as in [cons] U+200D U+09CD U+09AF. Following is a chart of consonants that are sometimes ligated with the Yaphala.

U+0996 U+09CD U+09AF

থ্য

U+0996 U+200D U+09CD U+09AF

থ্য

U+0997 U+09CD U+09AF

গ্য

U+0997 U+200D U+09CD U+09AF

গ্য

U+09A3 U+09CD U+09AF

ণ্য

U+09A3 U+200D U+09CD U+09AF

ণ্য

U+09A5 U+09CD U+09AF

ত্যা

U+09A5 U+200D U+09CD U+09AF

ত্যা

U+09A6 U+09CD U+09AF

দ্য

U+09A6 U+200D U+09CD U+09AF

দ্য

U+09A7 U+09CD U+09AF

ধ্য

U+09A7 U+200D U+09CD U+09AF

ধ্য

U+09A8 U+09CD U+09AF

ন্য

U+09A8 U+200D U+09CD U+09AF

ন্য

U+09AA U+09CD U+09AF

প্য

U+09AA U+200D U+09CD U+09AF

প্য

U+09AA U+09CD U+09AF
ব্য

U+09AE U+09CD U+09AF
ম্য

U+09AF U+09CD U+09AF
য্য

U+09B0 U+200C U+09CD U+09AF
র্য

U+09B2 U+09CD U+09AF
ল্য

U+09B6 U+09CD U+09AF
শ্য

U+09B7 U+09CD U+09AF
ষ্য

U+09B8 U+09CD U+09AF
স্য

U+09AC U+200D U+09CD U+09AF
ব্য

U+09AE U+200D U+09CD U+09AF
ম্য

U+09AF U+200D U+09CD U+09AF
য্য

U+09B0 U+200D U+09CD U+09AF
র্য

U+09B2 U+200D U+09CD U+09AF
ল্য

U+09B6 U+200D U+09CD U+09AF
শ্য

U+09B7 U+200D U+09CD U+09AF
ষ্য

U+09B8 U+200D U+09CD U+09AF
স্য

Yaphala After Independent Vowels

The yaphala is used after independent vowels. This is formed by typing [independent vowel] U+09CD U+09AF. Following are some examples.

Combination
U+0985 U+09CD U+09AF
অ ্ য = অ্য

U+0987 U+09CD U+09AF
ই ্ য = ই্য

U+0989 U+09CD U+09AF
উ ্ য = উ্য

U+098A U+09CD U+09AF
ঊ ্ য = ঊ্য

Example Word

অ্যাসিড

ই্যউ

উ্যেক্

ঊ্যৈড্‌জ্

U+098F U+09CD U+09AF

এ+্+য = এ়

এ্যালুমিনিয়াম

U+0993 U+09CD U+09AF

ও+্+য = ও়

ওয়েব

U+0994 U+09CD U+09AF

ঔ+্+য = ঔ়

ঔন্ট

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

U+0995 U+09CD U+09B0 U+0989

ক+্+র+উ=করুউ

U+0997 U+0989

গ+উ=গউ

U+09A6 U+09CD U+09B0 U+0989

দ+্+র+উ=দরুউ

U+09A7 U+09CD U+09B0 U+098A

ধ+্+র+উ=ধরুউ

U+09B6 U+0989

শ+উ=শউ

U+09B8 U+09CD U+09A4 U+0989

স+্+ত+উ=সতুউ

U+09B9 U+0989

হ+উ=হউ

U+09B9 U+098B

হ+ঋ=হঊ

Ligated Form

U+0995 U+09CD U+09B0 U+200D U+0989

ক+্+র+ZWJ+উ=করু

U+0997 U+200D U+0989

গ+ZWJ+উ=গু

U+09A6 U+09CD U+09B0 U+200D U+0989

দ+্+র+ZWJ+উ=দরু

U+09A7 U+09CD U+09B0 U+200D U+098A

ধ+্+র+ZWJ+উ=ধরু

U+09B6 U+200D U+0989

শ+ZWJ+উ=শু

U+09B8 U+09CD U+09A4 U+200D U+0989

স+্+ত+ZWJ+উ=স্তু

U+09B9 U+200D U+0989

হ+ZWJ+উ=হু

U+09B9 U+200D U+098B

হ+ZWJ+ঋ=হু

When typing Bengali script text, there are some consonant vowel combinations that users may desire to form an alternate shape. For example, the default is to place the Ukaar below the Sha. In the event a user would want the

Ukaar to display below in the alternate form, the ZWJ should be used to cause the joining behavior. The Bangla Academy has defined the following forms to behave as specified:

U+0997 + U+09C1

গ+়=গু

U+0997 U+200D U+09C1

গ+ZWJ+়=গু

U+09B6 + U+09C1

শ+়=শু

U+09B6 U+200D U+09C1

শ+ZWJ+়=শু

U+09B0 + U+09C1

র+়=রু

U+09B0 U+200D U+09C1

র+ZWJ+়=রু

U+09B0 + U+09C2

র+়=রু

U+09B0 U+200D U+09C2

র+ZWJ+়=রু

U+09B9 + U+09C1

হ+়=হু

U+09B9 U+200D U+09C1

হ+ZWJ+়=হু

Khanda Ta

The Unicode 4.0 preview makes a simplistic attempt at addressing the Khanda Ta. The documentation indicates that the Khanda Ta is the half form of the Ta (U+09A4).

U+09A4 U+09CD U+200D

ত+্+ZWJ → ঙ

Due to the nature of the Khanda Ta, it is necessary to use the ZWJ to make the half form and then implementers must not have the Khanda Ta considered as the pre-base consonant. The reason is that the following consonant should be able to function with matras. The following is an example of the problem because the Khanda Ta is considered as a half form of a consonant and is therefore automatically joined with the following consonant. The result is incorrect and does not exist.

U+09A4 U+09CD U+200D U+0995 U+09C7

ত+্+ZWJ+ক+ে → (ঙক

To avoid the behavior above and obtain the correct behavior, it is necessary to type a ZERO WIDTH NON-JOINER after the Khanda Ta to cause the cluster to break for correct placement of the matra. This necessity of using a ZWNJ to break the shaping should be an indicator that the Khanda Ta should be considered for its own code point if a different half form behavior for the Ta (U+09A4) can be defined.

U+09A4 U+09CD U+200D U+200C U+0995 U+09C7

ত+্+ZWJ+ZWNJ+ক+ে → ঙক

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.

U+09CD U+09AF

্+য → ্য

U+200D U+09CD U+09AF

ZWJ+্+য → ঙ

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.