Title:	**Use of ZWJ for Bengali CV "Ligatures"**
Doc. Type:	Expert contribution
Source:	Peter Constable, Microsoft
Date :	2006-02-06
Action:	For consideration by UTC
References:	
Distribution:	UTC members

Abstract

In Bengali, certain consonant + vowel mark combinations have two visual forms: one in which the vowel mark appears in its nominal form, and another in which a distinct conflated "ligature" typeform for the consonant-vowel combination is used. This document proposes a mechanism to distinguish these text elements in encoded representation using ZERO-WIDTH JOINER (ZWJ).

The problem

There are four Bengali consonant-vowel combinations for which there are two visual presentations: a "non-ligated" form in which the vowel mark takes its nominal form, and a "ligated" form in which the consonant and vowel combine into a distinct form:

Characters	Code points	Ligated	Non-ligated
ga + u-kaar	0997, 09C1		
ra + u-kaar	09B0, 09C1		
ra + uu-kaar	09B0, 09C2		
sha + u-kaar	09B6, 09C1		
ha + u-kaar	09B9, 09C1		
ha + rhi-kaar	09B9, 09C3		

The "ligated" forms are traditional forms. They are used in handwriting and some printing, but in modern printing the "non-ligated" forms are more common: they are used in newspapers and are associated with "modern" typefaces. (See Radice 1994, pp. 6, 11, 24, 43, 255, 270.) The traditional forms still appear to be preferred, however.

In Bengali text, no semantic distinction is made on the basis of these different presentations. At least some users consider it important that implementations support both, however, and that the distinction be representable in plain text.

The proposed solution

Whereas Bengali consonant conjuncts are formed using virama, virama is not appropriate in this case: the inherent vowel is not killed but is overridden by the vowel mark, and to introduce consonant + virama + vowel sequences would potentially destabilize the encoding model for Indic scripts.

Instead, these consonant-vowel conjoined forms can be treated as ligatures, and the general function of ZWJ and ZWNJ can be used for requesting or blocking the formation of ligatures. Thus, a given font implementation can choose whether or not to treat the "ligature" forms as defaults. If the non-ligated form is the default, then ZWJ can be used to request the ligature; for example:

Character sequence	Display
< 0997, 09C1 >	
< 0997, 200D, 09C1 >	

But if the ligated form is the default, then ZWNJ can be used to block the ligature:

Character sequence	Display
< 0997, 09C1 >	
< 0997, 200C, 09C1 >	

Possible concern: variation in the interaction between ZWJ and combining marks in combining mark sequences

One possible concern with this proposal is the variation in functionality of ZWJ in different contexts. In particular, the question arises as to how ZWJ should behave within combining mark sequences. The function of ZW(N)J in cursive connection has been documented in relation to Arabic for a long time, but there is no discussion of how presence of combining marks affects Arabic cursive connection when using ZW(N)J.

Some relevant earlier UTC precedents were set for Hebrew and Syloti Nagri scripts. For Hebrew, ZWJ can be used to "conjoin" the meteg with hataf vowels:

Character sequence	Display
< ALEF, HATAF PATAH, ZWJ, METEG >	

For Syloti Nagri, ZWJ can be used to "conjoin" a consonant with a following consonant when there is an intervening vowel mark, or to "conjoin" a vowel mark with a following consonant:

Character sequence	Display
< ব, ◌ী, ব, ZWJ, ◌ী, ৰ > ("bibir")	বীব্রী
< ক্ষ, ◌ী, ZWJ, ৰ > ("kir")	ক্ষীৰী

These precedents and the current proposal reflect a consistent principle: the ZWJ requests a ligature of the immediately preceding character, C1, with some following character, C2; C2 is either the immediately following combining mark or is the first base character following that combining-mark sequence. It is the choice of the following character that ligates with C1 that varies from one script to another according to the particular needs of that script.

A corollary of this proposal, therefore, is to propose the following principle regarding the ligation-forming effect of ZWJ in the context of combining mark sequences:

> Effect of ZWJ on ligature formation in combining mark sequences: The presence of ZWJ within a combining-mark sequence requests a ligature of the immediately preceding character with a following character. The following character may be the immediately following combining mark or the first base following that combining-mark sequence. Which following character is to become a component of the ligature is determined on a script-by-script basis.

References

Radice, William. 1994. *Teach yourself Bengali: a complete course for beginners.* Chicago: NTC Publishing Group.