L2/05-130

Title:	**Encoding Model for Syloti Nagri Conjoining Behaviour**
Doc. Type:	Expert contribution
Source:	Peter Constable, Microsoft
Date :	2005-05-07
Action:	For consideration by UTC
References:	L2/02-387, L2/02-388, L2/03-146r, L3/03-151r; UTC #93 minutes (L2/02-345); UTC consensus 95-C23
Distribution:	UTC members

Abstract

While Syloti Nagri is an Indic script, it is atypical of Indic scripts in relation to conjoining behaviour. In the initial proposal documents, an alternative to the virama encoding model was proposed. This was considered controversial, however, with various UTC members strongly opposing a model that was significantly different from existing virama models used for other Indic scripts. Accordingly, the proposal documents were revised to adopt a model like Burmese script, and the proposal was subsequently accepted.

In working on a rendering implementation for Syloti Nagri, however, we are finding that using the virama model alone is problematic for Syloti Nagri. We are requesting that Syloti Nagri conjoining behaviour be reviewed, and propose the use of virama where appropriate but other means for controlling ligation where the virama is not appropriate.

Historical review

In the initial proposal for Syloti Nagri script (L2/02-387 and L2/02-388), various issues related to conjoining behaviour in Syloti Nagri were discussed, and an alternative to the virama encoding model was proposed for Syloti Nagri. At UTC#93, the committee decided that this aspect of the proposal was controversial, and asked that a consensus among concerned parties be established:

> [93-A102] Action Item for Rick McGowan: Report back to Peter Constable that the model used in the Syloti Nagri proposal is controversial with respect to the use of ZWJ, and this issue needs to be resolved among concerned parties before the script can be accepted. The UTC recommends that existing rendering practice should be used if at all possible.

In relation to this action item, I produced document L2/03-146r comparing four alternative models. This was discussed in an ad-hoc committee consisting of Rick McGowan, Ken Whistler, Michael Everson and me. There was openness to two of the models, with overall preference given to one, the “Burmese” model.

Accordingly, the proposal was revised (L2/03-151r) to use the “Burmese” model:

> Our proposal has been revised to use an encoding model like that used for Myanmar script (model "D" in L2/03-146r): U+xx06 SYLOTI NAGRI HASANTA

> is used to determine formation of a conjunct of the preceding and following characters, in which case the HASANTA has no direct visual display. If, however, the HASANTA is followed by ZWNJ, no conjunct is formed, and the HASANTA becomes visible.

The revised proposal was approved by UTC#95:

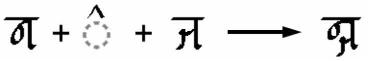
> [95-C23] Consensus: The UTC accepts the Syloti Nagri Script for encoding at A800..A82B with names as amended in discussion. Syloti Nagari Sign Ful (U+xx28 in the document) is renamed "Eight Pointed Asterisk", and is placed at U+2055. (Note: the name of the above character was changed by consensus 95-C27 to "Flower Punctuation Mark", as the glyph is not always 8-pointed.)

It should be noted that the UTC action did not make specific reference to an encoding model for Syloti Nagri conjoining behaviour, nor did the block description for Syloti Nagri included in Unicode 4.1 specify how conjoining behaviour should be controlled.

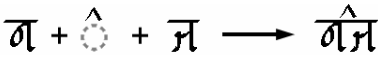
The outstanding problem

Document L2/03-146r compared various possible encoding models in terms of their ability to provide unique encoded representations for distinct Syloti Nagri text elements involving conjoined forms. There was one important factor was overlooked, however: the fact that that these ligature forms are discretionary, and as a result a typical font implementation will not support some ligatures that could conceivably be used. This is a problem with the “Burmese” encoding model since it can lead to display of an overt virama in places where that would be both inappropriate and incorrect. Yet, there is no way to inhibit the display of the virama.

To review, Syloti Nagri allows for conjoining of consonants in a consonant cluster — i.e. a consonant with the inherent vowel “killed” followed by another consonant. For example:



As for other Indic scripts, a rendering implementation is not absolutely required to display a conjoined form in such situations. If the conjoined form is not displayed, then the full form of both consonants is displayed along with the overt halant:



The halant (hasanta) is not commonly used — writers regularly write a vowel-less consonant with no explicit marking — but it is perfectly acceptable to display an overt halant in such a context.

The virama model is adequate and appropriate for representing Syloti-Nagri killed-vowel consonant conjuncts. Syloti Nagri is atypical of Indic scripts, however, in that it allows conjoining of many combinations of characters, not just consonant-consonant combinations. These can include:

- a vowel letter and a following consonant:

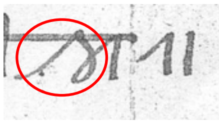


Figure 1: Vowel-consonant conjunct a-m in “amra” (‘we’)

- a vowel letter with anusvar and a following consonant:

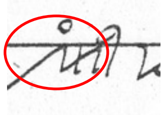


Figure 2: Vowel-consonant conjunct a-k with anusvara in “angki” (‘eye’)

- a vowel letter and a following vowel letter:

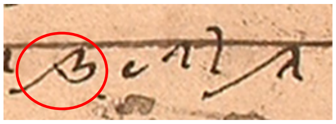


Figure 3: Vowel-vowel conjunct a-u in “auliar” (‘saint’s’)

- a spacing vowel sign and a following consonant:

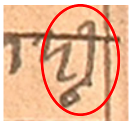


Figure 4: Conjunct dependent-í + n in “din” (‘day’)

Syloti Nagri texts have also included “false” conjunct forms: conjoined consonants representing the initial and final consonants of a live syllable (with vowel). These can include cases in which the vowel is the inherent vowel associated with the first consonant, or in which a explicit vowel mark occurs. Such “false” conjuncts are usually found at the ends of lines in manuscripts as a space-saving device. For example:

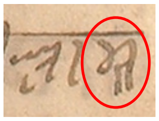


Figure 5: Line-ending “false” conjunct m-t in “alamot” (‘miracle’); vowel is the inherent vowel /o/

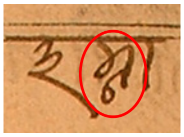


Figure 6: The word /iman/ (‘faith’), written with a “false” conjunct m-n; vowel is the vowel sign a

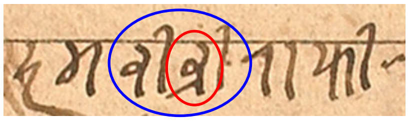


Figure 7: “False” conjunct b-r in “bibir” (‘lady’s’); vowel is the vowel sign i

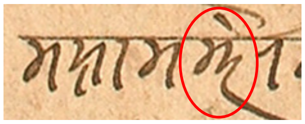


Figure 8: “False” conjunct m-d with e-kar in “mohammoder” (‘Muhammad’s’); vowel is the vowel sign e

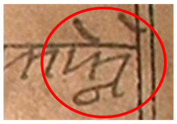


Figure 9: “False” conjunct k-n with two e-kar diacritics in “kene” (‘why’); vowel is the vowel sign e (with a second e for the following syllable)

These atypical conjoining forms are rare: they occur in old manuscripts and are not in current usage. They still need to be supported in Unicode implementations, however.

The “Burmese” virama-based encoding model is able to provide distinct encoded representations for most of these text elements. It cannot distinguish “false” conjuncts with spacing vowel signs, such as the example in Figure 7, from both “true” consonant conjuncts with a following vowel sign and also ligatures of a vowel-sign and a following consonant, such as the example in Figure 4: the same representation as one or the other of these must be used. In the original proposal, it was assumed that the same representation would be used as for “true” consonant conjuncts followed by a spacing vowel sign since there is no visual distinction: the single visual representation has two distinct readings.[1]

That encoding model does *not* deal well, however, with the fact that these conjoined forms are discretionary ligatures that may only be supported in specialty fonts designed to represent ancient manuscripts. In particular, the encoding model requires each of these conjoined forms to be represented with a virama, but in none of these cases would it be appropriate to display the text with an overt halant.

The following table summarizes the display results for each of these cases under different rendering conditions: (a) using specialty fonts for paleographic applications, (b) the incorrect display that this encoding would produce with typical fonts designed for modern usage that do not support rare conjoining forms, and (c) the correct display results that should be obtained when using non-specialty fonts.

[1] A reader can determine that a “false conjunct” reading, such as /bir/ is correct rather than a “true conjunct” reading, such as /bri/, based on the context. For further details, see the discussion of “false” conjuncts in L2/03-388 (p. 22ff).

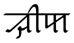
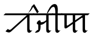
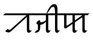
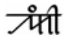
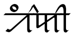
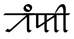
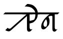
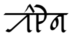
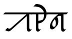
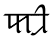
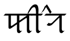
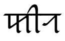
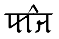
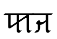
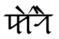
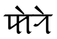
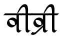
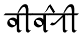
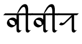
Case	Example sequence	Conjoining display with specialty font	Bad non-conjoining display with basic font	Correct non-conjoining display
V + C	< ꠀ, virama, ꠔ, ◌ꠤ, ꠇ > ("atiko")			
V + anusvara + C	< ꠀ, ◌ꠋ, virama, ꠇ, ◌ꠤ > ("angki")			
V + V	< ꠀ, virama, ꠄ, ꠘ > ("aeno")			
V-sign + C	< ꠇ, ◌ꠤ, virama, ꠞ > ("kir")			
"false" conjunct" C + C	< ꠇ , virama, ꠔ > ("kot")			
"false conjunct" with non-spacing mark	< ꠇ, ◌ꠦ, virama, ꠞ, ◌ꠦ > ("kere")			
"false conjunct" with spacing vowel mark	< ꠛ, ◌ꠤ, ꠛ, virama, ꠞ, ◌ꠤ > ("bibir" — same sequence as "bibri")			

Table 1: Comparison of display results for atypical conjoinable sequences under different display environments assuming "Burmese" virama-based encoding model

Consider the following scenario, then: a researcher has a Web site reproducing traditional Sylheti poetry found in a collection of ancient manuscripts. She encodes rare conjoining forms that were used in the original manuscripts. Some readers visiting the site have fonts intended for modern usage, however. When they view the text, several words appear with a virama in places that make the text unreadable and meaningless. Consequently, they submit bug reports to the owner of the site. This is a problem resulting directly from the use of a virama to control conjoining behaviour in contexts in which a halant mark is inappropriate.

Therefore, we find that there are two sets of cases in which conjoining behaviour occurs:

- The most common cases, "true" consonant conjuncts involving consonant clusters ("dead" consonant + consonant), for which it is acceptable to display a halant mark.
- Rare cases involving a vowel letter, vowel sign or live (with-vowel) consonant conjoining with a following letter in which it is *un*acceptable to display a halant mark.

The use of a virama to control conjoining in the first set of cases is consistent with the encoding model of Indic scripts, and is reasonable and unproblematic. For the latter cases, however, the use of a virama is not consistent with other Indic scripts since the conjoining behaviour is atypical of Indic scripts, and the use of a virama leads to significant display problems.

As a result, I propose that these two sets of cases be distinguished in terms of how conjoining is handled:

- Keep the virama model for handling "true" consonant conjuncts.
- The other cases involve discretionary ligation; therefore, handle them like other cases of discretionary ligation: ligatures can be enabled using font mechanisms such as OpenType features; if there is a need to request a ligature explicitly in encoded representation, this can be done using the existing function of ZERO WIDTH JOINER defined for this purpose.

The use of ZWJ avoids the display problems shown in Table 1. It would also make it possible to distinguish "false" conjuncts with a spacing vowel mark from other cases. The display results that would be obtained using ZWJ for the various cases using different fonts are summarized in Table 2:

Case	Example sequence	Conjoining display with specialty font	Conjoining display with basic font
V + C	< ꠀ, ZWJ, ꠔ, ◌ꠤ, ꠇ > ("atiko")	ꠀ‍ꠔꠤꠇ	ꠀ‍ꠔꠤꠇ
V + anusvara + C	< ꠀ, ZWJ, ◌ꠋ, ꠇ, ◌ꠤ > ("angki")	ꠀ‍ꠋꠇꠤ	ꠀ‍ꠋꠇꠤ
V + V	< ꠀ, ZWJ, ꠄ, ꠘ > ("aeno")	ꠀ‍ꠄꠘ	ꠀ‍ꠄꠘ
V-sign + C	< ꠇ, ◌ꠤ, ZWJ, ꠞ > ("kir")	ꠇꠤ‍ꠞ	ꠇꠤ‍ꠞ
"false" conjunct" C + C	< ꠇ , ZWJ, ꠔ > ("kot")	ꠇ‍ꠔ	ꠇ‍ꠔ
"false conjunct" with non-spacing mark	< ꠇ, ZWJ, ◌ꠦ, ꠞ, ◌ꠦ > ("kere")	ꠇ‍ꠦꠞꠦ	ꠇ‍ꠦꠞꠦ
"false conjunct" with spacing vowel mark	< ꠛ, ◌ꠤ, ꠛ, ZWJ, ◌ꠤ, ꠞ > ("bibir" – distinct sequence from "bibri")	ꠛꠤꠛ‍ꠤꠞ	ꠛꠤꠛ‍ꠤꠞ

Table 2: Comparison of display results for atypical conjoinable sequences using ZWJ

This would make use of the Indic virama model where appropriate, but not where it isn't. It uses an already-defined mechanism for controlling discretionary ligation that avoids the display of a halant mark in inappropriate places. It uses the typical Indic encoding model for the common conjoining cases, and leaves the special use of ZWJ for the rare cases. This combination of encoding mechanisms presents no issues for implementation, and provides for display requirements for Syloti Nagri whereas the virama model alone cannot.