L2/03-146R

Doc. Type:	Working Group document
Title:	Alternate encoding models for Syloti Nagri
Source:	Peter Constable, SIL International
Status:	Individual contribution
Action:	For consideration by UTC and JTC 1/SC 2/WG 2
Date:	2003-5-5

Alternate Encoding Models for Syloti Nagri

Peter Constable, SIL International

1: Introduction

In L2/02-388, two possible encoding models for Syloti Nagri were considered: the standard “virama” encoding model typically used for Indic scripts, and an alternate that was considered in that document preferable for various reasons. In spite of those reasons, there has been some concern at what may be perceived as a new encoding model, even though the mechanisms used all exist already in Unicode (though not all are used with some Indic script).

This document summarises those two models, and also compares two other possible models that have been suggested since L2/02-388.

In this document, the following conventions are used:

Co = live consonant with inherent vowel

Cd = dead consonant

Vn = independent vowel

Vs = spacing dependent vowel (a-kar, i-kar)

Vc = combining vowel mark (e-kar, u-kar)

# = word boundary.

2: Four encoding models

The four encoding alternatives will be identified as “A”, “B”, “C” and “D”. Briefly, the four models are the following:

A. standard virama model as presented in L2/02-388: Dead consonants are indicated by virama, and virama is used to indicate conjunct formation. If a conjunct is to be prevented, this is done by placing ZWJ or ZWNJ after virama. If ZWNJ is used, overt hasanta appears; if ZWJ is used, there is no overt marking of the dead consonant.

B. ZWJ-based model proposed in L2/02-388: Conjunct formation is controlled by using ZWJ between the characters to be ligated. Overt hasanta is represented as a distinct combining mark, independent of any control of conjunct formation. When there is no overt hasanta, a distinction between a live consonant with inherent vowel and a dead consonant can be encoded using an invisible character SYLOTI NAGRI INHERENT VOWEL.

C. virama model with distinct hasant character: Conjunct formation is controlled by a character SYLOTI NAGRI VIRAMA. Overt hasanta is represented as a distinct combining mark, independent of any control of conjunct formation. When there is no overt hasanta, a distinction between a live consonant with inherent vowel and a dead consonant can be encoded using an invisible character SYLOTI NAGRI INHERENT VOWEL.

D. Burmese virama model: Conjunct formation is controlled by a character SYLOTI NAGRI VIRAMA. The virama used alone results in a conjunct (except, of course, if word-final); virama followed by ZWNJ results in an overt hasanta rather than a conjunct. When there is no overt hasanta, a distinction between a live consonant with inherent vowel and a dead consonant can be encoded using an invisible character SYLOTI NAGRI INHERENT VOWEL.

Under model C, the only thing that would be particularly unusual for Syloti Nagri in comparison to more familiar Indic scripts is that the encoded character “virama” would potentially occur after all kinds of characters—consonants, independent vowels (spacing and non-spacing), dependent vowels, anusvara—rather than just consonants.

3: Comparison of alternative models

The four alternate models will be compared in relation to several different types of written form that need to be supported.

It is assumed here that situations involving a dead consonant are be distinguished from situations involving a live consonant with inherent vowel for processing where absence of vowel must be distinguish from inherent vowel. In common usage, though, the inherent vowel character (“IV”) in models B, C and D would not normally be used.

3.1 Situation: Co# or CoC

In common usage where the distinction between dead consonants and live consonants with inherent vowel is not encoded, all four models would use the same encoded representation, as shown for model A.

Sample word = ꠇꠔ “koto”.

Model	Encoded Representation	Example
A	< …C, #… >, < …C, C… >	< ꠇ, ꠔ >
B	< …C, IV, #… >, < …C, IV, C… >	< ꠇ, IV, ꠔ, IV >
C	< …C, IV, #… >, < …C, IV, C… >	< ꠇ, IV, ꠔ, IV >
D	< …C, IV, #… >, < …C, IV, C… >	< ꠇ, IV, ꠔ, IV >

3.2 Situation: Cd# or CdC

Sample word = फाागजफा “kantok”.

Model	Encoded Representation	Example
A	< …C, virama, ZWJ, #… >, < …C, virama, ZWJ, C… >	< फा, ◌ा, ग, virama, ZWJ, ज, फा, virama, ZWJ >
B	< …C, #… >, < …C, C… >	< फा, ◌ा, ग, ज, फा, >
C	< …C, #… >, < …C, C… >	< फा, ◌ा, ग, ज, फा, >
D	< …C, #… >, < …C, C… >	< फा, ◌ा, ग, ज, फा, >

3.3 Situation: Cd + overt hasanta + #, Cd + overt hasanta + C

Sample word = फााग̂जफा̂ “kantok”.

Model	Encoded Representation	Example
A	< …C, virama, ZWNJ, #… >, < …C, virama, ZWNJ, C… >	< फा, ◌ा, ग, virama, ZWNJ, ज, फा, virama, ZWNJ >
B	< …C, hasanta, #… >, < …C, hasanta, C… >	< फा, ◌ा, ग, hasanta, ज, फा, hasanta >
C	< …C, hasanta, #… >, < …C, hasanta, C… >	< फा, ◌ा, ग, hasanta, ज, फा, hasanta >
D	< …C, virama, ZWNJ, #… >, < …C, virama, ZWNJ, C… >	< फा, ◌ा, ग, virama, ZWNJ, ज, फा, virama, ZWNJ >

3.4 Situation: Cd + C conjunct

Sample word = फााग्जफा “kantoko”.

Model	Encoded Representation	Example
A	< …C, virama, C… >	< फा, ◌ा, ग, virama, ज, फा >
B	< …C, ZWJ, C… >	< फा, ◌ा, ग, ZWJ, ज, फा >
C	< …C, virama, C… >	< फा, ◌ा, ग, virama, ज, फा >
D	< …C, virama, C… >	< फा, ◌ा, ग, virama, ज, फा >

3.5 Situation: Vn + C conjunct

Sample word = ꠀ꠆ꠔꠤꠇ “atiko”.

Model	Encoded Representation	Example
A	< …Vn, virama, C… >	< ꠀ, virama, ꠔ, ◌ꠤ, ꠇ >
B	< …Vn, ZWJ, C… >	< ꠀ, ZWJ, ꠔ, ◌ꠤ, ꠇ >
C	< …Vn, virama, C… >	< ꠀ, virama, ꠔ, ◌ꠤ, ꠇ >
D	< …Vn, virama, C… >	< ꠀ, virama, ꠔ, ◌ꠤ, ꠇ >

3.6 Situation: Vn + Vn conjunct

Sample word = ꠀ꠆ꠄꠘ “aeno”.

Model	Encoded Representation	Example
A	< …Vn, virama, Vn… >	< ꠀ, virama, ꠄ, ꠘ >
B	< …Vn, ZWJ, Vn… >	< ꠀ, ZWJ, ꠄ, ꠘ >
C	< …Vn, virama, Vn… >	< ꠀ, virama, ꠄ, ꠘ >
D	< …Vn, virama, Vn… >	< ꠀ, virama, ꠄ, ꠘ >

3.7 Situation: Vn-anusvara + C conjunct

Sample word = ꠀꠋ꠆ꠇꠤ “angki”.

Model	Encoded Representation	Example
A	< …C, anusvara, virama, C… >	< ꠀ, ◌ꠋ , virama, ꠇ, ◌ꠤ >
B	< …C, anusvara, ZWJ, C… >	< ꠀ, ◌ꠋ , ZWJ, ꠇ, ◌ꠤ >
C	< …C, anusvara, virama, C… >	< ꠀ, ◌ꠋ , virama, ꠇ, ◌ꠤ >
D	< …C, anusvara, virama, C… >	< ꠀ, ◌ꠋ , virama, ꠇ, ◌ꠤ >

3.8 Situation: Vs + C conjunct

Sample word = ꠇꠤ꠆ꠞ “kir”.

Model	Encoded Representation	Example
A	< …Vs, virama, C… >	< ꠇ, ◌ꠤ, virama, ꠞ >
B	< …Vs, ZWJ, C… >	< ꠇ, ◌ꠤ, ZWJ, ꠞ >
C	< …Vs, virama, C… >	< ꠇ, ◌ꠤ, virama, ꠞ >
D	< …Vs, virama, C… >	< ꠇ, ◌ꠤ, virama, ꠞ >

3.9 Situation: C-Vc + C false conjunct

Sample word = प्रै “kere”.

Model	Encoded Representation	Example
A	< …C, Vc, virama, C… >	< पा, ◌े, virama, ग, ◌े >
B	< …C, Vc, ZWJ, C… >	< पा, ◌े, ZWJ, ग, ◌े >
C	< …C, Vc, virama, C… >	< पा, ◌े, virama, ग, ◌े >
D	< …C, Vc, virama, C… >	< पा, ◌े, virama, ग, ◌े >

3.10 Situation: C-Vs + C false conjunct

This situation is completely indistinguishable in all four models from that of a Cd + C conjunct followed by Vs (see §0).

3.11 Situation: Co + C false conjunct

This situation cannot be distinguished under model A from that of a Cd + C conjunct (see §0). In common usage where the distinction between dead consonants and live consonants with inherent vowel is not encoded, models B, C and D would likewise use the same representation as for Cd + C conjucts. The significant point here is that these models make a distinction possible, while model A does not.

Sample word = प्र “kot”.

Model	Encoded Representation	Example
A	*(same as Cd + C conjunct)*	
B	< …C, IV, ZWJ, C… >	< पा , IV, ZWJ, ज >
C	< …C, IV, virama, C… >	< पा , IV, virama, ज >
D	< …C, IV, virama, C… >	< पा , IV, virama, ज >

4: Summary

Apart from the use of ZWJ in model B versus a virama character in the other models, there is no real difference between these models for several of the situations considered:

- Cd + C conjunct (see §0)
- Vn + C conjunct (see §0)
- Vn + Vn conjunct (see §0)
- Vn-anusvara + C conjunct (see §3.7)
- Vs + C conjunct (see §0)
- C-Vc + C false conjunct (see §0)
- C-Vs + C false conjunct (none of the models are able to distinguís this from Cd + C conjuncts; see §0)

For some of the situations considered, the models are differentiated only in special usage contexts in which a distinction between dead consonants and live consonants with inherent vowel needs to be representable. In common usage contexts, in which this distinction is not normally encoded, the models are comparable. The situations in question are the following:

- Co# or CoC (in common usage, not distinguished from Cd# or CdC; see §3.1)
- Co + C false conjunct (in common usage, not distinguished from Cd + C conjunct; see §3.11)

In both of these situations, it is model A that is significantly different from the others: it has a significantly different means than the other models of distinguishing Co# or CoC from Cd# or CdC, and it is not capable of distinguishing Co + C false conjuncts from Cd + C conjuncts.

The key situations in which the models are differentiated are two:

- Cd# or CdC (see §0)
- Cd + overt hasanta + #, Cd + overt hasanta + C (see §3.3)

In the former situation, it is again model A that is significantly different from the others. In the latter situation, models A and D are the same, and models B and C are the same.