Encoding model to represent conjuncts in Syloti Nagri

Srinidhi A
srinidhi.pinkpetals24@gmail.com

Sridatta A
sridatta.jamadagni@gmail.com

November 29, 2017

1 Introduction

This document analyses the model to represent conjuncts in Syloti Nagri and changes required in Indic Syllabic category of A806 SYLOTI NAGRI SIGN HASANTA.

2 Background

In 2005 based action item 103-C14 it appears that Virama model is adopted.

[103-C14] *Consensus:* *The UTC accepts the proposal on Syloti Nagri conjoining behavior as documented in L2/05-130 such that: "... use virama where appropriate but using other means for controlling ligation where the virama is not appropriate."*

Virama is also known as Hasanta in Bengali and Sylheti languages.

L2/05-130 recommends following

- Virama model for representing true conjuncts.

- For rare handwritten ligatures involving vowels and false conjuncts, other means like use of OpenType features and ZWJ.

While Virama model is adequate and unproblematic, the second recommendation involving cross-cluster ligatures and false conjuncts needs further analysis and examination.

3 Rare cross cluster ligatures and false conjuncts

It is observed that these occurrences are highly rare and occur in some handwritten sources. These are not attested modern printed sources as indicated in the above document.

3.1 Cross cluster ligatures involving vowels and vowel signs

In handwritten manuscripts having cursive nature of various scripts, the letters are often joined at informal level. These ligatures may not require distinction to represent separately in plaintext involving ZWJ. We believe use of ZWJ could complicate the model and could be handled at font level using OpenType features.
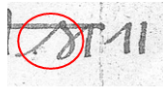
○ a vowel letter and a following consonant:

Figure 1: Vowel-consonant conjunct a-m in "amra" ('we')
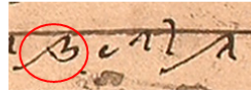
○ a vowel letter with anusvar and a following consonant:

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')
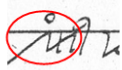
○ a spacing vowel sign and a following consonant:

Figure 4: Conjunct dependent-í + n in "din" ('day')

Figure 1. Cross cluster ligatures as provided in L2/05-130.

3.2 False conjuncts

As said in the L2/05-130 *"Such "false" conjuncts are usually found at the ends of lines in manuscripts as a space-saving device".*
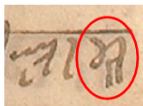
Figure 5: Line-ending "false" conjunct m-t in "alamot" ('miracle'); vowel is the inherent vowel /o/
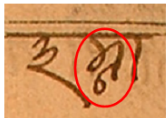
Figure 6: The word /iman/ ('faith'), written with a "false" conjunct m-n; vowel is the vowel sign a
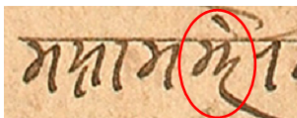

Figure 8: "False" conjunct m-d with e-kar in "mohammoder" ('Muhammad's'); vowel is the vowel sign e
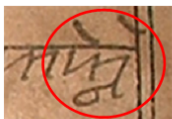

Figure 9: "False" conjunct k-n with two e-kar diacritics in "kene" ('why'); vowel is the vowel sign e (with a second e for the following syllable)

Figure 2. False conjuncts as provided in L2/05-130.

These false conjuncts may be termed as scribal errors and practices. In order to represent these formations following sequences were provided in L2/05-130.

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	< ㅈ, ZWJ, न, ी, पा > ("atiko")	ऋीपा	ऋनीपा
V + anusvara + C	< ㅈ, ZWJ, ं, पा, ी > ("angki")	नंंी	नंंपी
V + V	< ㅈ, ZWJ, पे, न > ("aeno")	ऋेन	ऋपेन
V-sign + C	< पा, ी, ZWJ, न > ("kir")	पाी	पाीन
"false" conjunct" C + C	< पा , ZWJ, न > ("kot")	प्न	पान
"false conjunct" with non-spacing mark	< पा, ZWJ, ं, न, ं > ("kere")	प्ने	पोने
"false conjunct" with spacing vowel mark	< व, ी, व, ZWJ, ी, न > ("bibir" — distinct sequence from "bibri")	वीव्री	वीवीन

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

Here for conjunct प्न (kta), ZWJ is used to identical to Virama, this sequence C1+ZWJ+C2 is not found in Indic scripts to form conjuncts.

However, for bibir both व+ ी+ व +ZWJ+ ी+ न and व+ ी+ व +HASANTA+ न+ी sequences produce same text वीव्री ,which produces visual ambiguity and hence creating problems in security.

Unicode encodes written forms based on orthography not on pronunciation or reading. To simplify encoding model these false conjuncts should be represented based on how they are written using virama regardless of their reading of the text.

This is similar to the model adopted in Malayalam handling ഠഠ.

As a consequence the graphical sequence ഠഠ in text is ambiguous in reading. The reader must generally use the context to understand if ഠഠ is read /rara/ or /tta/. It is only when a vowel part appears between the two ഠ that the reading cannot be /tta/. Note that similar situations are common in many other orthographies. For example, *th* in English can be a digraph (*cathode*) or two separate letters (*cathouse*); *gn* in French can be a digraph (*oignon*) or two separate letters (*gnome*).

The sequence <0D31, 0D31> is rendered as ഠഠ, regardless of the reading of that text. The sequence <0D31, 0D4D, 0D31> is rendered as ഩ. In both cases, vowels signs can be used as appropriate, as shown in *Table 12-35*.

4 Editorial changes required in Core Specification

The following description is provided on virama and conjuncts in Chapter 15.1.

*"**Virama and Conjuncts.** Syloti Nagri is atypical of Indic scripts in use of the virama (hasanta) and conjuncts. Conjuncts are not strictly correlated with the phonology being represented. They are neither necessary in contexts involving a dead consonant, nor are they limited to such contexts. Hasanta was only recently introduced into the script and is used only in limited contexts. Conjuncts are not limited to sequences involving dead consonants but can be formed from pairs of characters of almost any type (consonant, independent vowel, dependent vowel) and can represent a wide variety of syllables. It is generally unnecessary to overtly indicate dead consonants with a conjunct or explicit hasanta. The only restriction is that an overtly rendered hasanta cannot occur in connection with the first element of a conjunct. The absence of hasanta does not imply a live consonant and has no bearing on the occurrence of conjuncts. Similarly, the absence of a conjunct does not imply a live consonant and has no bearing on the occurrence of hasanta."*

- This text is confusing and we find Syloti Nagri is not atypical in Indic scripts.
- The Sylheti language orthography is similar to Hindi, Bengali and other Indo-Aryan languages having medial and final schwa deletion like "*The absence of hasanta does not imply a live consonant and has no bearing on the occurrence of conjuncts*" and "*the absence of a conjunct does not imply a live consonant and has no bearing on the occurrence of hasanta*".
- The text may be revised and updated with text similar to Devanagari or Bengali.

5 Conclusion

- It is requested to discuss and review on the cross-cluster ligatures and false conjuncts if at all they require distinct representation in plain-text and document the appropriate model in Core specification, otherwise model proposed in this document can be adopted.
- Update description on Virama and Conjuncts in Core specification accordingly.
- A806 SYLOTI NAGRI SIGN HASANTA is currently assigned Indic_Syllabic_Category=Pure_Killer. Pure Killer refers to killing of inherent vowel in consonant sequence, with no consonant stacking behavior. As Hasanta is used as both visible killer viramas and consonant stackers, the property should be changed to Indic_Syllabic_Category=Virama.