Visual Ambiguity Involving Indic Consonant Letters

Peter Constable
October 9, 2017

Summary

Unicode Chapter 12 provides explicit guidance that vowel letters in text should be represented as atomic characters and not as visually-similar character sequences involving a different vowel letter plus a vowel mark (matra). This document describes a different class of sequences that result in visually-similar representations for consonant letters, and proposes that additional guidance be added to Chapter 12 recommending that consonant letters be represented only as atomic characters.

Existing guidance involving vowel letters

Many vowel letters in Indic scripts have an appearance similar to a different vowel letter with a vowel mark added. This creates the potential for different encoded representations to represent the same text element. This can result in problems for interoperability, as well as security risks. Clearly only one encoded representation is intended and desired. To that end Chapter 12 provides this guidance:

> ***Vowel Letters.*** Vowel letters are encoded atomically in Unicode, even if they can be analyzed visually as consisting of multiple parts. *Table 12-1* shows the letters that can be analyzed, the single code point that should be used to represent them in text, and the sequence of code points resulting from analysis that should not be used.

Table 12-1. Devanagari Vowel Letters

For	Use	Do Not Use
ऍ	0904	<0905, 0946>
आ	0906	<0905, 093E>
ई	0908	<0930, 094D, 0907>
ऊ	090A	<0909, 0941>
ऎ	090D	<090E, 0945>

Visual ambiguity involving consonant letters

Chapter 12, in the section *Explicit Half Consoants*, specifies the coded representation for consonant half forms:

> "To explicitly encode a half-consonant form, the Unicode Standard adopts the convention of placing the character U+200D ZERO WIDTH JOINER immediately after the encoded dead consonant."

It also notes that this encoded representation does not require a base consonant letter:

> This encoding of half-consonant forms also applies in the absence of a base letterform. That is, this technique may be used to encode independent half-forms, as shown in *Figure 12-5*.

Figure 12-5. Independent Half-Forms in Devanagari

$GA_d + ZWJ \rightarrow GA_h$

ग् + [ZWJ] → ग

Note that the visual appearance of many Indic consonant letters includes a vertical stem, which is also has the appearance of the dependent vowel mark AA. This is seen in the following Devanagari examples — the red, ovelaid element is the form of U+093E DEVANAGARI VOWEL SIGN AA.:

Also note that many half forms have the appearance of the consonant letter with the vertical stem removed:

Table 12-2. Sample Devanagari Half-Forms

क	+	◌्	+	ZWJ	→	क्	न	+	◌्	+	ZWJ	→	न्
ख	+	◌्	+	ZWJ	→	ख्	प	+	◌्	+	ZWJ	→	प्
ग	+	◌्	+	ZWJ	→	ग्	फ	+	◌्	+	ZWJ	→	फ्
घ	+	◌्	+	ZWJ	→	घ्	ब	+	◌्	+	ZWJ	→	ब्
च	+	◌्	+	ZWJ	→	च्	भ	+	◌्	+	ZWJ	→	भ्
ज	+	◌्	+	ZWJ	→	ज्	म	+	◌्	+	ZWJ	→	म्
झ	+	◌्	+	ZWJ	→	झ्	य	+	◌्	+	ZWJ	→	य्
ञ	+	◌्	+	ZWJ	→	ञ्	ल	+	◌्	+	ZWJ	→	ल्
ण	+	◌्	+	ZWJ	→	ण्	व	+	◌्	+	ZWJ	→	व्
त	+	◌्	+	ZWJ	→	त्	श	+	◌्	+	ZWJ	→	श्
थ	+	◌्	+	ZWJ	→	थ्	ष	+	◌्	+	ZWJ	→	ष्
ध	+	◌्	+	ZWJ	→	ध्	स	+	◌्	+	ZWJ	→	स्

The examples shown here involve single consonants, though the same apply in the case of some consonant conjunct forms, such as ksha:

Consonant conjunct *ksha*: क्ष

Half consonant *ksha*: क्

These three points are significant:

- The vertical stem found in many consonants has the same appearance as vowel sign aa.
- Many half consonant forms have the appearance of the consonant letter with the vertical stem stem removed.
- Explicit half consonant forms have an encoded representation.

These facts create a potential for visual ambiguity: the visual form of a consonant can potentially be represented either as a consonant letter or as a sequence using the encoded representation of an explicit half consonant followed by the vowel sign aa.

Such alternate representations are known to be in use. For example, Microsoft has recently received a bug report regarding a Web page (http://www.ipr.res.in/hindiconf15/documents/committees.html) with a character sequence using the alternate representation with a half form that is expected by the authors to be displayed as a full consonant form:

Author-expected display as *ksha*:

Character sequence used in page:
< 0915 KA, 094D VIRAMA, 0937 SSA, 094D VIRAMA, 200D ZWJ, 093E VOWEL SIGN AA>

The character sequence for *ksha* assumed in Unicode is < 0915 KA, 094D VIRAMA, 0937 SSA>.

Note that some application support display of such sequences as shown, though others will treat this as an invalid sequence — an ill-formed cluster due to having a vowel sign without a valid base consonant:

Display as ill-formed cluster:

These alternate encoded representations are visually indistinct from the simpler encoded representations assumed by Unicode, as atomic characters or basic conjunct sequences. There is no useful semantic distinction that can be made if the visual forms are identical, hence no valid user scenario to require a plain-text distinction. These alternate sequences only add potential for problems in interoperability or security. Their use should be deprecated.

For this reason, it is proposed that the text of Chapter 12 be extended to provide explicit guidance against use of this class of sequences, just as has been done for vowel letters.

Scope of impact

The issue described here affects at least the north Indic scripts, Devanagari, Bengali, and Gujarati. In Devanagari, it affects all consonants that have a right-side vertical stem, and consonant conjuncts for which the nominal form has a right-side vertical stem. Cases for Gujarati will be similar to Devanagari. Bengali script has fewer consonants that have half forms.

The issue likely also affects Newa script, covered in Chapter 13, and may also affect Kaithi, Siddham and Tirhuta scripts, covered in Chapter 15. Additional text in those chapters may also be warranted.