**L2/07-043**

# Public Review Issue #96

***Revision 2 01-10-2006*** *Clarified the text, added figures, fixed some typos, and broadened slightly the use of ZWJ for half-consonants.*

*To submit feedback, please see http://www.unicode.org/review/*

## Allowing Joiner Characters in Identifiers

*This PRI affects the use of ZWJ and ZWNJ in identifiers and may be relevant in a variety of contexts, including such areas as international domain names for Arabic, Persian, and languages of India such as Hindi and Malayalam.*

The use of format characters in identifiers is problematical because the formatting effects they represent are considered merely stylistic or otherwise out of scope for identifiers. To make matters worse, it's possible to misapply format characters such that users can create strings that look the same but actually contain different characters.

For these reasons format characters are normally excluded from Unicode identifiers. However, visible distinctions created by certain format characters (particularly the *joiner controls*) are necessary and carry meaning in certain languages. A blanket exclusion of format characters makes it impossible to create identifiers based on certain words or phrases in those languages. Identifier systems that attempt to provide more natural representations of terms, such as geographic names, company names, and so on should consider allowing these characters, but limited to particular contexts where they are necessary.

The goal for such a restriction of format characters to particular contexts is to

a. allow the use of these characters where required in normal text
b. exclude as many cases as possible where no visible distinction results
c. be simple enough to be easily implemented with standard mechanisms such as regular expressions

Normal usage, as meant here, does not include technical usage such as mathematical expressions or pedagogical use (eg, illustration of half-forms or joining forms in isolation).

*We would especially appreciate feedback as to:*

1. Whether we can restrict the list of Viramas?
   - In particular, in which scripts of South East Asia would the addition of ZWNJ or ZWJ after a Halant (Virama) not result in visually-different strings?
   - That is, which scripts do not have visually-different conjuncts with Halant (Virama), and do not have specialized forms like chillu? (Thai is one example.)
2. Are there any other contexts requiring these characters?
3. Can the contexts below be made simpler? For example, in practical use, do we need to allow for marks after the virama in the cases below?
4. The conjunct contexts are described in terms of Letters, since that is the simplest formulation in terms of the current Unicode properties. Should we narrow it to be only consonants (in the relevant scripts)?
   - Since there is no Unicode property for virama-consonants, it would require either an explicit list or a new property. The characters it would be limited to are listed under a subheading "Consonants" in the charts for the various scripts, such as in http://www.unicode.org/charts/PDF/U0900.pdf.

## Proposal

Allow joiner controls (U+200C ZERO WIDTH NON-JOINER [ZWNJ] and U+200D ZERO WIDTH JOINER [ZWJ]) in the Unicode recommendations for identifiers, in limited contexts as specified below. In each of the following cases, the context listed must also only consist of characters from a single script (after ignoring *Common* and *Inherited* script characters). Note that since the contexts are limited to being in the same script, the conjunct contexts are only applicable to scripts containing a Virama *(halant)* character.

Parsing identifiers can be a performance-sensitive task. However, these characters are quite rare in practice, thus the regular expressions (or equivalent processing) only rarely would need to be invoked. Thus these tests should not add any significant performance cost overall.

The characters and their contexts are given by the following:

**A. ZWNJ in the following contexts:**

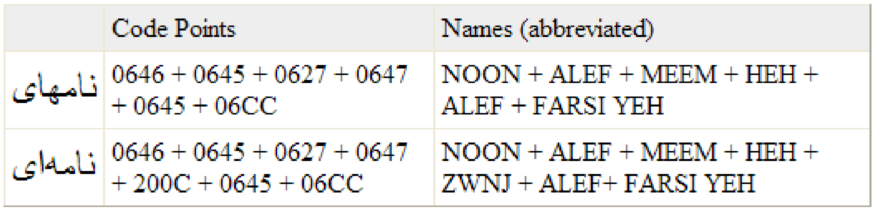
1. **Breaking a cursive connection.** That is, in the context based on the Arabic Shaping property, consisting of:
   - A Left-Joining character, followed by zero or more Transparent characters, followed by a ZWNJ, followed by zero or more Transparent characters, followed by a Right-Joining character
   - This corresponds to the following regular expression (in Perl-style syntax):

     /$L $T* ZWNJ $T* $R/
     where:

     - $T = [:Joining_Type=Transparent:]
     - $R = [[:Joining_Type=Dual_Joining:][: Joining_Type=Right_Joining:]]
     - $L = [[:Joining_Type=Dual_Joining:][:Joining_Type=Left_Joining:]]

   - **Example:** Farsi <Noon, Alef, Meem, Heh, Alef, Farsi Yeh>. Without a ZWNJ, it translates to "names"; with a ZWNJ between Heh and Alef, it means "a letter". Figure 1 illustrates this.

## Figure 1.



| | Code Points | Names (abbreviated) |
|---|---|---|
| نامهای | 0646 + 0645 + 0627 + 0647 + 0645 + 06CC | NOON + ALEF + MEEM + HEH + ALEF + FARSI YEH |
| نامه‌ای | 0646 + 0645 + 0627 + 0647 + 200C + 0645 + 06CC | NOON + ALEF + MEEM + HEH + ZWNJ + ALEF+ FARSI YEH |

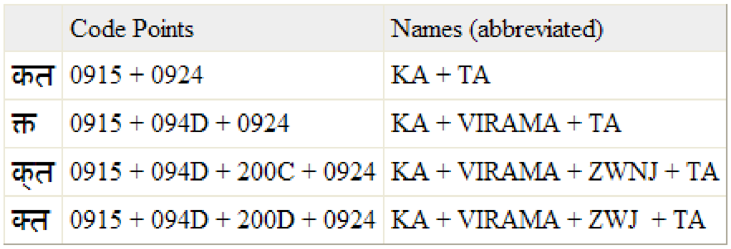
2. **In a conjunct context.** That is, a sequence of the form:
   - A Letter, followed by zero or more combining marks, followed by a Virama, followed by a ZWNJ, followed by zero or more combining marks, followed by an Letter.
   - This corresponds to the following regular expression (in Perl-style syntax):

     /$L $M* $V $M* ZWNJ $M* $L/
     where:

     - $L = [:General_Category=Letter:]
     - $M = [:General_Category=Mark:]
     - $V = [:Canonical_Combining_Class=Virama:]

   - **Example:** In Malayalam, ZWJ and ZWNJ are used to make distinctions involving cillu forms. (See p. 337 of TUS 5.0.) The status would change if and when the cillu forms are separately encoded in Unicode.
   - **Example:** Figure 2 shows the use of ZWJ and ZWNJ in Devanagari.

## Figure 2.



| | Code Points | Names (abbreviated) |
|---|---|---|
| कत | 0915 + 0924 | KA + TA |
| क्त | 0915 + 094D + 0924 | KA + VIRAMA + TA |
| क्‌त | 0915 + 094D + 200C + 0924 | KA + VIRAMA + ZWNJ + TA |
| क्‍त | 0915 + 094D + 200D + 0924 | KA + VIRAMA + ZWJ + TA |

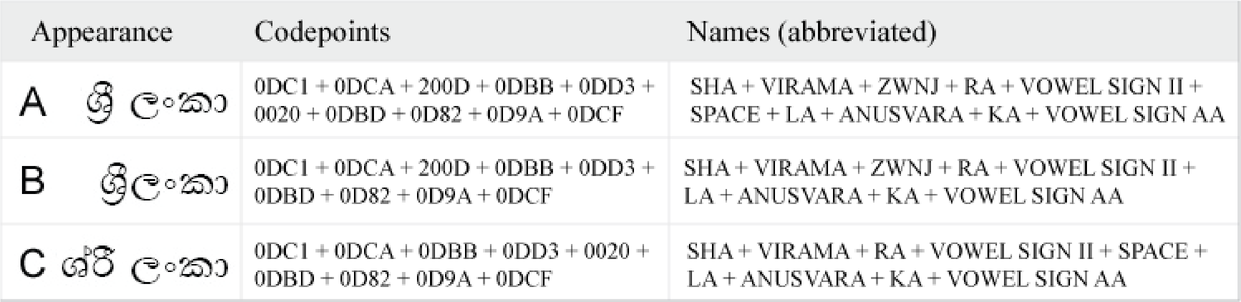
**B. ZWJ in the following contexts:**

1. **In a conjunct context.** That is, a sequence of the form:
   - A Letter, followed by zero or more combining marks, followed by a Virama, followed by zero or more combining marks, followed by a ZWJ.
   - This corresponds to the following regular expression (in Perl-style syntax):

     /$L $M* $V $M* ZWJ/
     where:

     - $L = [:General_Category=Letter:]
     - $M = [:General_Category=Mark:]
     - $V = [:Canonical_Combining_Class=Virama:]
   - **Example:** Devanagari RA + VIRAMA + ZWJ + KA (see also Figure 2)
   - **Example:** The Sinhala word for the country 'Sri Lanka' in Figure 3A, which uses both a space character and a ZWJ. Removing the space gives the text in Figure 3B which is still readable, but removing the ZWJ completely modifies the appearance of the 'Sri' cluster and gives the text in Figure 3C.

**Figure 3.**



| Appearance | Codepoints | Names (abbreviated) |
|---|---|---|
| A ශ්‍රී ලංකා | 0DC1 + 0DCA + 200D + 0DBB + 0DD3 + 0020 + 0DBD + 0D82 + 0D9A + 0DCF | SHA + VIRAMA + ZWNJ + RA + VOWEL SIGN II + SPACE + LA + ANUSVARA + KA + VOWEL SIGN AA |
| B ශ්‍රීලංකා | 0DC1 + 0DCA + 200D + 0DBB + 0DD3 + 0DBD + 0D82 + 0D9A + 0DCF | SHA + VIRAMA + ZWNJ + RA + VOWEL SIGN II + LA + ANUSVARA + KA + VOWEL SIGN AA |
| C ශ්රී ලංකා | 0DC1 + 0DCA + 0DBB + 0DD3 + 0020 + 0DBD + 0D82 + 0D9A + 0DCF | SHA + VIRAMA + RA + VOWEL SIGN II + SPACE + LA + ANUSVARA + KA + VOWEL SIGN AA |

*Note: To provide for canonically equivalent orderings of combining marks, the optional combining marks ($M) and transparent characters ($T) are allowed in multiple locations.*