

L2/07-146

Public Review Issue #96

Allowing Special Characters in Identifiers

Revision 3 04-19-2007 *Significantly tightened the requirements for ZWJ and ZWNJ by reducing the number of possible scripts, and simplifying the sequences. Also added the equivalent characters needed for Mongolian.*

This PRI affects the use of special characters (ZWJ, ZWNJ and Mongolian separators) in identifiers. It may be relevant in a variety of contexts, including such areas as international domain names for Arabic, Persian, Sinhalese, Khmer, and Malayalam. If you believe that there are any other languages requiring the use of special characters, please respond as directed on <http://unicode.org/review> and include the PRI number and Revision Number in your message.

The use of format characters in identifiers is problematical because the formatting effects they represent are normally just 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, which can create security problems (see [UTR# 36: Unicode Security Considerations](#)).

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).

Proposal

Allow joiner controls (U+200C ZERO WIDTH NON-JOINER [ZWNJ] and U+200D ZERO WIDTH JOINER [ZWJ]) and Mongolian separators (U+202F NARROW NO-BREAK SPACE [NNBSP] and U+180B .. U+180D *mongolian free variation selectors*) in the Unicode recommendations for identifiers, but only in very limited contexts as specified below.

Script Restriction. In each of the following cases, the specified sequence must only consist of characters from a single script (after ignoring *Common* and *Inherited* script characters).

Performance. 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 a Virama, followed by a ZWNJ, followed by an Letter, where the Letters and Virama are all in the *Malayalam* script, or they are all in the *Khmer* script
 - *Issue: is the Khmer inclusion required semantically?*
- This corresponds to the following regular expression (in Perl-style syntax): `/L $V ZWNJ L/` where:
 - \$L = [[:General_Category=Letter:]]
 - \$V = [[:Canonical_Combining_Class=Virama:]]
- **Example:** In Khmer, U+17A2 U+200D(ZWNJ) U+17CA U+17B7 U+17A2 U+17BB U+17CA U+17C7 [អ័អ័ៈ] is a case where the first TRIISAP needs to be escaped, but the second does not (as there is a below base vowel).
- **Example:** The Malayalam word for *eyewitness*. The form without the ZWNJ is incorrect in this case.

Figure 2.

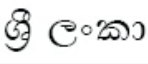
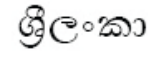
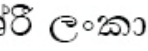
	Code Points	Names (abbreviated)
ദക്ഷിണാക്ഷി	0D28 + 0D43 + 0D15 + 0D4D + 200C + 0D38 + 0D3E + 0D15 + 0D4D + 0D37	DA + SIGN VOCALIC R + KA + VIRAMA + ZWNJ + SA + SIGN AA + KA + VIRAMA + SSA
ദക്ഷി	0D28 + 0D43 + 0D15 + 0D4D + 0D38 + 0D3E + 0D15 + 0D4D + 0D37	DA + SIGN VOCALIC R + KA + VIRAMA + SA + SIGN AA + KA + VIRAMA + SSA

B. ZWJ in the following context:

1. In a conjunct context. That is, a sequence of the form:

- A Letter, followed by a Virama, followed by a ZWJ, where the Letter and Virama are both in the Sinhala script
- This corresponds to the following regular expression (in Perl-style syntax): `/L $V ZWJ/` where:
 - \$L = [[:General_Category=Letter:]]
 - \$V = [[:Canonical_Combining_Class=Virama:]]
- **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 	0DC1 + 0DCA + 200D + 0DBB + 0DD3 + 0020 + 0DBD + 0D82 + 0D9A + 0DCF	SHA + VIRAMA + ZWJ + RA + VOWEL SIGN II + SPACE + LA + ANUSVARA + KA + VOWEL SIGN AA
B 	0DC1 + 0DCA + 200D + 0DBB + 0DD3 + 0DBD + 0D82 + 0D9A + 0DCF	SHA + VIRAMA + ZWJ + RA + VOWEL SIGN II + LA + ANUSVARA + KA + VOWEL SIGN AA
C 	0DC1 + 0DCA + 0DBB + 0DD3 + 0020 + 0DBD + 0D82 + 0D9A + 0DCF	SHA + VIRAMA + RA + VOWEL SIGN II + SPACE + LA + ANUSVARA + KA + VOWEL SIGN AA

C. Mongolian Separators (NNBSP or MVSs) in the following context:

- Between Mongolian Letters.** That is, a sequence of the form:
 - o A Mongolian Letter, followed by NNBSP or a MVS, followed by a Mongolian Letter.
 - o This corresponds to the following regular expression (in Perl-style syntax): `/$ML $MS $ML/` where:
 - `$ML = [[:General_Category=Letter:]&[:Script=Mongolian:]]`
 - `$MS = [\u202F \u180B \u180C \u180D]`
 - o **Example:** See pages 454 455 of *The Unicode Standard, Version 5.0*.

Comparison Cases

The above description restricts the usage of Joiner and Nonjoiner quite substantially from Revision 1 of this Public Review Issue. This restriction was based on a review of the cases where these characters would be required for semantic differences relevant to identifiers. The other specified cases of Joiner or Nonjoiner usage in the Unicode Standard were not considered to be required for identifiers. They are listed here for comparison, so that reviewers can look over these cases to see if there are good reasons for including them in the above list.

Non-Semantic

Cases that do not carry semantic differences (or at least differences which are not sufficient to be required in identifiers for modern languages):

1. Devanagari, Half-forms as in Tables 9-2 and 9-4, pp 309, 311
2. Bengali, Figures 9-10 and 9-11, p 314; and RA + JOINER + VIRAMA + YA, p 316
3. Gurmukhi Table 9-10, p 320
4. Kannada, p 334
5. Myanmar, p 380
6. Buginese, Figure 11-5, p 398
7. Phags-Pa, Table 10-2 (since Phags-Pa is a historic script, it is not suitable for general purpose identifiers).
8. Sinhala, use of ZWJ in front of the virama to form touching consonants, "used in classical and Buddhist texts".

Superseded

Sequences that have been superseded in usage by other characters, or should be in the near future (the characters having already been approved by the Unicode consortium, and slated for Unicode 5.1):

1. Devanagari, RA + VIRAMA + ZWJ
2. Bengali, TA + VIRAMA + ZWJ
3. Myanmar, LETTER + VIRAMA + ZWNJ (see [UTN #11](#))
4. Malayalam, LETTER + VIRAMA + ZWJ