L2/07-399

To: Unicode Technical Committee

From: Peter Edberg, Apple Inc.

Date: Oct. 18, 2007

Re: Proposed bugfix/enhancement for UAX #29

Consider the following examples using “sentence continuation” or “mid-sentence” punctuation (comma, colon, semicolon, em-dash) following an ambiguous full stop:

- When in the U.S., Tom lives in Ohio.
- “In the U.S.,” Mary said.
- It helps (in the U.S.), Mary said.
- In the U.S.: Tom will help you.
- “In the U.S.”: Aux Etats-Unis.
- The three groupings are Canada and the U.S.; France, Germany, and Denmark; and Italy, Spain, and Greece.
- In the U.K.—Scotland excluded—tennis is popular.

In each case, applying UAX #29 currently results in incorrectly finding a sentence break before the first continuation punctuation following “U.S.” or “U.K.”; this does NOT happen if the first letter after the continuation punctuation is lowercase instead of uppercase, because of rule SB8 in UAX #29.

There are other situations in which UAX #29 specifies an incorrect sentence break, as shown by the following examples (incorrect break is after “U.S.”):

- 私はU.S.にゆきました。
- U.S.→Mexico→Panamaという順番で旅行をしました。

These other situations are not addressed by this proposal (and in fact the second example may be beyond the scope of what UAX #29 should address). The problems in the first set of examples (involving “sentence continuation” punctuation) can be addressed as follows:

1. Define a new sentence break property value, say SContinue, and add an entry for it in Table 4:

SContinue	Any of the following characters: U+002C COMMA U+3001 IDEOGRAPHIC COMMA U+FE10 PRESENTATION FORM FOR VERTICAL COMMA U+FE11 PRESENTATION FORM FOR VERTICAL IDEOGRAPHIC COMMA U+FF0C FULLWIDTH COMMA U+003A COLON U+FE13 PRESENTATION FORM FOR VERTICAL COLON U+FF1A FULLWIDTH COLON U+003B SEMICOLON U+FE14 PRESENTATION FORM FOR VERTICAL SEMICOLON U+FF1B FULLWIDTH SEMICOLON U+2014 EM DASH U+FE31 PRESENTATION FORM FOR VERTICAL EM DASH U+002D HYPHEN-MINUS U+FF0D FULLWIDTH HYPHEN-MINUS

2. Change rule SB8a as follows (new text underlined)**:**

SB8a. (STerm | ATerm) Close* Sp* × (SContinue | STerm | ATerm)

3. Change the text before SB6 accordingly, for example (new text underlined)**:**

Do not break after ambiguous terminators like period, if they are immediately followed by a number or lowercase letter, if they are between uppercase letters, ~~or~~ if the first following letter (optionally after certain punctuation) is lowercase, or if they are followed by “continuation” punctuation such as comma, colon, or semicolon. For example, a period may be an abbreviation or numeric period, and thus may not mark the end of a sentence.