Segmentation of Regional Indicator Symbols

Authors: Markus Scherer & Mark Davis & Andy Heninger
Date: 2012-August-02
Live version: http://goo.gl/kjvMN

Regarding PRI #212 (line break, UAX #14) & PRI #215 (grapheme & word breaks, UAX #29)

Note: Regional indicator symbols are U+1F1E6..U+1F1FF REGIONAL INDICATOR SYMBOL LETTER A..REGIONAL INDICATOR SYMBOL LETTER Z.

Proposal

We propose changing UAX #14 & #29 so that RI characters join together, and require some other character to separate them. Basically these are the rules:

RI × RI
÷ RI
RI ÷

Note: Other than review notes, the current draft updates of UAX #14 & #29 contain no changes other than for handling of regional indicator symbols.

Use the Zero Width Space as a separator

We propose changing line break properties of regional indicator letters so that they normally stick together, and recommend using U+200B ZWSP as a separator if necessary. This is different than the previously suggested approaches of using ZWJ (or CGJ) to prevent line breaks in the middle of pairs of regional indicator symbols.

UAX #14 Line breaks

1. Add one new lb property value, RI=Regional_Indicator.
2. Assign the 26 regional indicator symbols lb=RI. (In Unicode 6.1 they have lb=AL=Alphabetic.)
3. Add a new rule LB30a. *Don't break between regional indicators*
    - RI × RI

UAX #29 Grapheme cluster breaks

Starting from the U6.1 version of UAX #29:
1. Add one new GCB property value, RI=Regional_Indicator.
2. Assign the 26 regional indicator symbols GCB=RI. (In Unicode 6.1 they have GCB=XX=Other.)
3. Add a new rule GCB8a. *Don't break between regional indicators*
    - RI × RI

UAX #29 Word breaks

Starting from the U6.1 version of UAX #29:

1. Add one new WB property value: RI=Regional_Indicator.
2. Assign the 26 regional indicator symbols WB=RI. (In Unicode 6.1 they have WB=XX=Other.)
3. Add a new rule WB13c. *Don’t break between regional indicators*
    - RI × RI

Rationale

The current draft update to UAX #14, which moves the Zero Width Joiner out of lb=CM, is disruptive to a complex and fragile set of rules. In particular, splitting class CM is problematic because rule LB9 causes CM* to be ignored in following rules; moving some characters out of CM requires the new class to be added to several of the following rules, which has not been done completely in the current draft update, and it is difficult to ensure that all of the edge cases are covered.

We originally proposed using U+034F CGJ which has lb=GL and thus has the desired effect on lb=ID=Ideographic characters, and requires no change to the set of lb=CM characters, and no change to the UAX #14 rule set.

However, in UTC discussion there was concern that requiring the insertion of a glue character is fragile and inconvenient. An approach of separating sequences of regional indicator letters was favored.

Also there was the feeling that for single flags (the most common case), the above approach is simpler than using CGJ (while if multiple adjacent flag characters were most common, CGJ might be better).