Segmentation of Regional Indicator Symbols

Authors: Markus Scherer & 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 as follows, based on the Unicode 6.1 versions, instead of the current draft updates posted for Unicode 6.2. Other than review notes, the current draft updates contain no changes other than for handling of regional indicator symbols.

Use the COMBINING GRAPHEME JOINER

We propose using the sequence <regional indicator symbol, U+034F CGJ, regional indicator symbol> (rather than a sequence with U+200D ZWJ) to prevent line breaks in the middle of pairs of regional indicator symbols.

UAX #14 Line breaks

1. No change to UAX #14 compared to Unicode 6.1.
2. Assign the 26 regional indicator symbols lb=ID, which provides the desired behavior in combination with CGJ. (In the Unicode 6.2 beta, the regional indicator symbols already have lb=ID.)

UAX #29 Grapheme cluster breaks

Starting from the current draft update of UAX #29:

1. Add two new GCB property values, RI=Regional_Indicator and GL=Glue, rather than After_Joiner and Joiner.
2. Assign U+034F CGJ the value GCB=GL.
3. Assign the 26 regional indicator symbols GCB=RI. (In Unicode 6.1 they have GCB=XX=Other.)
4. Modify rules GB8b and GB9 to use RI & GL rather than AJ & J.

UAX #29 Word breaks

Starting from the current draft update of UAX #29:

1. Add two new WB property values, RI=Regional_Indicator and GL=Glue, rather than After_Joiner and Joiner.
5. Assign U+034F CGJ the value WB=GL.
2. Assign the 26 regional indicator symbols WB=RI. (In Unicode 6.1 they have WB=XX=Other.)
3. Modify rules WB3c and WB4 to use RI & GL rather than AJ & J.

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.

Our proposal takes advantage of the fact that CGJ's lb=GL already has the desired effect on lb=ID=Ideographic characters, and thus requires no change to the set of lb=CM characters, and no change to the UAX #14 rule set.

For grapheme cluster breaks and word breaks, our proposal is the same as in the current draft updates, except for the following:

1. We are using CGJ as the glue character, instead of ZWJ.
2. We are proposing property value names of "Regional_Indicator" instead of "After_Joiner" and "Glue" instead of "Joiner".