JTC 1/SC 2/WG 2 N3779

Title: Proposal on use of ZERO WIDTH JOINER (ZWJ) between two Regional Indicator Symbols

Date: 2010-3-3

Source: Japan

Status: NB Position

Requested Action: WG2 to discuss and to update 10646

Proposal on use of ZERO WIDTH JOINER (ZWJ)
between two Regional Indicator Symbols

1. Background

In its Tokyo meeting, WG2 decided to allocate 26 special symbols in UCS whose intended purpose is to represent some *emoji* symbols that are associated with some regional entities that we can identify with two-letter codes. As a result, WG2 N3738 (10646/FPDAM 8 ballot text) contains the set of 26 Regional Indicator Symbols.

An example scenario presented at the Tokyo meeting was as follows: When a transcoder received a *K-emoji* symbol F7 90 for a symbol of a region identified as US, it emits a sequence of two UCS characters: REGIONAL INDICATOR SYMBOL LETTER U (1F1FA) followed by REGIONAL INDICATOR SYMBOL LETTER S (1F1F8). In reverse, the transcoder maps a two-character sequence 1F1FA 1F1F8 back to K-emoji symbol F7 90.

2. What's wrong with it?

We need two Regional Indicator Symbols to represent an *emoji* symbol for a regional entity. A Regional Indicator Symbol has no clue in itself whether it is used as the first of two or the second. For example, REGIONAL INDICATOR SYMBOL LETTER U may be used as the first of US or as the second of RU.

Suppose you have a database system that stores *emoji* texts received from mobile phones and tried to find a particular text that contains an *emoji* symbol for a regional entity identified as US. What should the database system do when it sees a sequence of three characters: REGIONAL INDICATOR SYMBOL LETTER R, REGIONAL INDICATOR SYMBOL LETTER U, and REGIONAL INDICATOR SYMBOL LETTER S? The database system has no clue whether the frist two, R and U, represents a *K-emoji* F349 or the last two, U and S, represents a *K-emoji* F790, as long as it sees the UCS text locally. It needs to scan the text from the beginning to the end, in sequence, to know the position of the INDICATOR SYMBOL LETTER U, to determine which pairing was correct. Transcoders, text searches, text drawers, and other software that handle UCS

encoded *emoji* will live a tough life under the current scheme.

3. Proposed solution

In this contribution the author proposes use of ZERO WIDTH JOINER (ZWJ: 200D) between the two Regional Indicator Symbols that intend to represent a single regional symbol.

In our previous example, the same fragment of the database text is encoded as "... R ZWJ U S ..." if the R and U represents an *emoji* symbol, and as "... R U ZWJ S ..." if the U and S represents an *emoji* symbol. The text search software can simply look for a sequence of "U ZWJ S" to find an *emoji* symbol for a regional entity identified by US.

The character ZERO WIDTH JOINER is chosen in a text drawing (rendering) application in mind; A process of rendering a pair of Regional Indicator Symbols into a single glyph can be seen as a change of the graphic symbols of the two Regional Indicator Symbols to join to form a special shape; the first one to join to the second, and the second one to join to the first. So it seems natural to put a ZWJ between the two.

4. Proposed changes to the draft amendment

In a sentence between the name list for Regional Indicator Symbols, a phrase something like: "in a pairs joined with ZERO WIDTH JOINER" should be used to explain the intended use of the characters.

In EmojiSrc.txt file, put ZERO WIDTH JOINER between two Regional Indicator Symbols to show the source reference, i.e.,

```
1F1E8 200D 1F1F3;;F3D2;FBB3
1F1E9 200D 1F1EA;;F3CF;FBAE
1F1EA 200D 1F1F8;;F348;FBB1
1F1EB 200D 1F1F7;;F3CE;FBAD
1F1EE 200D 1F1F9;;F3D0;FBAF
1F1EF 200D 1F1F5;;F6A5;FBAB
```

```
1F1F0 200D 1F1F7;;F3D3;FBB4
1F1F7 200D 1F1FA;;F349;FBB2
1F1FA 200D 1F1F0;;F3D1;FBB0
1F1FA 200D 1F1F8;;F790;FBAC
```

(END OF DOCUMENT)