L2/170-434

Re: **Emoji Regex & UTS #51 Definitions**
From: **Mark Davis**
Date: **2017-12-22**

The Definitions that we have in UTS #51 are fairly complex. Yet in the end we still depend on the RGI test to filter the set of emoji sequences. Because of that dependency on RGI, we could simplify the definitions to make a simpler EBNF and Regex for detecting possible emoji characters and sequences. This has advantages for parsing and scanning text for emoji. We may also want to simplify the UTS #51 definitions.

Proposal

1. Add a section to the UTS #51 that describes the usage of the following EBNF and Regex expressions in scanning text to identify possible emoji.
2. Discuss in the UTC whether we would want to pursue simplifying our UTS #51 definitions to correspond more to the EBNF.

Issue

The following EBNF can be used to scan for *possible* emoji, which can then be verified by looking at the RGI sequences in the emoji data files. For brevity here, it uses some proposed property aliases, such as EMD for Emoji_Modifier. It is much simpler than the expressions currently in Definitions: you can think of it as the convex hull of what is matched by those definitions. Of course, it also includes some degenerate cases as a by-product of that simplicity, but those are weeded out by the RGI verification in any event.

EBNF	**Notes**
`possible_emoji := possible_zwj_element (\x{200D} possible_zwj_element)+`	200D = joiner
`possible_zwj_element :=`	RI = Regional_Indicator
`  \p{RI} \p{RI}`	
`| \p{Emoji} emoji_modification?`	
`emoji_modification :=`	Mn= Nonspacing_Mark
`  \p{EMD}`	FE0F = emoji VS
`| \x{FE0F}? \p{Mn}*`	E00xx are tags
`| [\x{E0020}-\x{E007E}]+ \x{E007F}`	E007F is the TERM tag.

From this a regex can be generated, as below. While it may seem complex, it is far simpler than what would result from the Definitions, which result in regex expressions which are many times more complicated, and yet still require verification using the list of RGI sequences.

Regex
`(\p{RI} \p{RI}`
`| \p{Emoji} (\p{EMD} | \x{FE0F}? \p{Mn}* | [\x{E0020}-\x{E007E}]+ \x{E007F})?)`
`   ( \x{200D}`
`(\p{RI} \p{RI}`
`| \p{Emoji} (\p{EMD} | \x{FE0F}? \p{Mn}* | [\x{E0020}-\x{E007E}]+ \x{E007F})?))+`

Notes

Note that this **EBNF** shares a characteristic with the UTS #51 Definitions: it is finer-grained than a grapheme cluster. That is, if you have “A<zwj><emoji>”, the Unicode 10 grapheme cluster rule GB11 will cause that to be one grapheme cluster, while a scanner based on the **EBNF** will find an emoji *within* that grapheme cluster. (Similarly, a scanner looking for ASCII letters will find an “A” within that same grapheme cluster.)

The proposed Unicode 11.0 GB11 rule is tighter, since it requires an emoji (or emoji-like) character before the ZWJ. So a sequence like “A<zwj><emoji>” will not count as a grapheme cluster under that new rule.

In any event, a subsequent more precise test for RGI sequences would also “split” a grapheme cluster like SUSHI+ZWJ+PLANE+ZWJ+SPONGE, by finding it to contain 3 emoji (separated by ZWJ characters). It is a known feature of the stock grapheme cluster rules that they are not built to handle degenerate cases if that makes the rules too complicated.