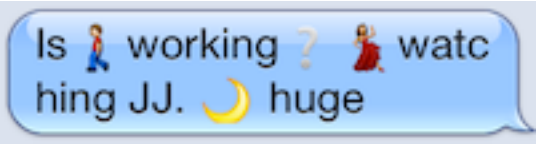
L2/12-198

Date: 2012-May-11
To: UTC
From: Peter Edberg
Subject: Line break issues with characters used for emoji

Emoji are now being used in many contexts other than Japanese e-mail and text messages. Because the current line break property value for most Unicode characters used for emoji is AL (alphabetic letter), this is causing problems in many of these contexts. For example, here is a text message that mixes Latin characters with emoji; the first line consists of just Latin letters and emoji, with no spaces, and is breaking in an inappropriate place (this would be less of an issue in Japanese, because most Japanese characters have line break property value ID and could break before or after emoji):



This problem can occur in any language (including Japanese) with long strings of emoji, which are treated as a single word. Such strings are also becoming popular as part of an Internet trend of trying to write funny sentences completely in emoji.

While there seems to be a significant general problem with the line break property values for symbols, there is an acute problem specifically with the characters added in Unicode 6.0 for representation of emoji. In advance of an overall review of line break property values for symbols (which I also think is necessary), I would like to suggest a PRI for the following more limited proposal:

For Unicode characters added in 6.0 that are listed in EmojiSources.txt, have General Category "So" and currently have LineBreak class AL, change the LineBreak class to ID.

(The restriction to newly added characters is specifically to avoid having to consider backward compatibility issues with pre-6.0 characters unified with emoji). This would cover all of the emoji characters in the first line of the sample message above, and would address most of the line break issues for emoji.