L2/18-173

Re: Line Break changes for U11.0

From: Mark Davis

Date: 2018-05-02

The proposed update to UAX #14 for Unicode 11.0.0 (draft 1) contains a proposed change to rule LB8a. In earlier discussion during the UTC meeting we decided to delay those, and push to CLDR. However, I think we should revise that decision.

Emoji-Related Line Break rules as of Unicode 10.0

> ***LB8a*** *Do not break between a zero width joiner and an ideograph, emoji base or emoji modifier.*
>
> ZWJ × (ID | EB | EM)

> ***LB22*** *Do not break between two ellipses, or between letters, numbers or exclamations and ellipsis.*
>
> …
>
> (ID | EB | EM) × IN
>
> …

> ***LB23a*** *Do not break between numeric prefixes and ideographs, or between ideographs and numeric postfixes.*
>
> PR × (ID | EB | EM)
>
> (ID | EB | EM) × PO

> ***LB30a Break between two regional indicator symbols if and only if there are an even number of regional indicators preceding the position of the break.***
>
> *sot (RI RI)* RI × RI*
>
> *[^RI] (RI RI)* RI × RI*

> ***LB30b*** *Do not break between an emoji base and an emoji modifier.*
>
> EB × EM

Here is the rationale for the above rules in Unicode 10.0. The goal was to make minimal changes to Line Break to allow for LB8a, LB30a, and LB30b: that is, not break emoji ZWJ sequences, flag sequences, or emoji modifier sequences. The easiest way to do that at the time was to utilize ID, which includes most emoji characters. So there was a change in Unicode 9.0 to add emoji characters were to ID (where they didn't impact other rules) and

then to carve out EB and EM out of ID for use in LB30b, and thus change LB22 and LBB23a to include the union..

It was recognized at the time that ID was only an approximation, but it covered all the then-current emoji sequences. *But it doesn't any longer*.

Changes in Unicode 11.0.0 (draft 1)

[http://www.unicode.org/reports/tr14/proposed.html#LB8a]

> ***LB8a*** *Do not break between a **zero width joiner** and an **ideograph**, **emoji base** or **emoji modifier**.*
>
> ZWJ × ~~(ID | EB | EM)~~[\p{extended_pictographic}-\p{block=ASCII}-\p{RI}]
>
> This rule prevents breaks within most emoji zwj sequences, as defined by *ED-16. emoji zwj sequence* in [UTS51].
>
> Review Note: The change in 8a is to address a problem found in testing: the changes in UAX #29 for Extended Pictographic were not extended to Line Break (CLDR and ICU have been customized to use Extended Pictographic since 2016). The removal of the keycap emoji (the ASCII characters) and RI characters are because they don't occur in emoji ZWJ sequences. The change adds the same future-proofing & missing emoji as in UAX #29, CLDR, and ICU, and removes a large number of ideographs that were only included in 8a because it originally had made the changes for emoji simpler. This change has not, however, been reviewed yet by the UTC, and feedback is especially welcome.
>
> ~~Further customization of this rule may be necessary for best behavior of emoji zwj sequences, using [CLDR].~~
>
> [http://www.unicode.org/reports/tr14/proposed.html#Customization]
>
> ~~*Note:* Some changes to rules and data are needed for the best segmentation behavior of emoji zwj sequences [UTS51]. Implementations are strongly encouraged to use the the line break rules in the latest version of CLDR (Version 31 or later) [CLDR] and the latest emoji properties (version 5.0 or later) [UTS51].~~

During the current UTC meeting, the point was brought up that other rules should be changed, rules that also contain ID, EB, and EM. However, on further examination, it appears that the rationale for changing LB8a in Unicode 11 does not apply to the other emoji-related rules, and they should not be changed; see below. Moreover, the change does solve a real problem for Unicode 11.0.

Rationale for changes to LB8a in Unicode 11.0

1. As with changes in UAX 29, the use of \p{extended_pictographic} in LB8a helps to "future proof" for future emoji. But in UAX 14, the use of \p{extended_pictographic} also prevents breakage in ZWJ sequences that include Emoji characters that are outside of ID. That is, it covers emoji that ID was only an approximation for. Those

are (for Unicode 10.0) the following **372 code points**:

[☠❣💤💢▲⛩⚓☆⚡❄☄✨♠♥♦♣🎼♪🎶☎⌨💱💲✏✒⚒⚔⚙⚗⚖🔗⚰⚱⚠
🔞☢☣⬆↗➡↘⬇↙⬅↖↕↔↩↪⤴⤵🔃🔄🔙-🔝⚛🕉✡☸☯✝☦☪☮♈-♓⛎🔀-🔂
▶⏩⏭⏯◀⏪⏮🔼⏫🔽⏬⏸-⏺⏏🔅🔆♀♂⚕♻⚜⭕✅☑✔✖❌❎➕-➗➰➿✳✴
❇‼⁉❓-❕❗©®™🔟💯🔠-🔤🅰🆎🅱🆑-🆓ℹ🆔Ⓜ🆕🆖🅾🆗🅿🆘-🆚▪▫◻-◾⬛⬜
🔶-🔻💠🔘🔲🔳⚪⚫⬤🔴🔵]

There is one character in particular that will cause problems: Emoji 11 includes the pirate flag, which will break under the current formulation of Line Break.

2. The use of \p{extended_pictographic} also excludes most ideographs from the ZWJ behavior in rule LB8a. That includes the following **169,435 Code Points**. While these are degenerate cases (we don't much care about sequences of ZWJ followed by these), it is cleaner to exclude them from the ZWJ rule LB8a.

[⺀-⻿⼀-⿟⿰-⿻〄〆〇〒〓〔-〟〠-〿〿あいうえお-ぢつ-もやゆよ-ろわ-ゔゟアイウエオ-ヂツ-モヤユヨ-ロワ-ヴヷ-ヺㄅ-ㄯ㄰-ㆎ㆐-㆟ㆠ-ㆿ㇀-㇯㈀-㈞㈠-㉇㉐
㉑-㊿㋀-㋾㌀-㏿䷀-䷿一-鿿ꀀ-꓆꤀-꥟ꥠ-꥿가-힯ힰ-퟿豈-﫿︐-︙︰-︴﹅-﹆﹐﹑﹒﹔-﹗﹟﹠-﹡﹣-﹦﹨﹫＃＆＇＊＋－／
０-９＜-＞＠-Ｚ＼＾-ｚ｜～ｦｧ-ﾝﾞ-ﾟﾠ-ￜ￠-￢￣-￦-
ウ-𛀀-𛄞🈀🈁-🈂🈐-🈻🉀-🉈🉐🉑-]

Rationale for no other changes in Unicode 11.0

1. There is no need to change (ID | EB | EM) in anything but LB8a. (This was a question raised in the UTC.) Those lines were originally ID, and the use of (ID | EB | EM) is the normal process for a partition "carve out". There is no need (or desire) to include additional emoji in those rules.
2. In the future, some of the emoji that were added to ID could be moved out, but that is not necessary for Unicode 11.0.