Re: **Broadening ZWJ segmentation support**
To: **UTC**
From: **Mark Davis, Emoji SC**
Date: **2016-05-05**

It looks like Windows is adding more ZWJ sequences (ninja-cat-the-windows-only-emoji/), with additional characters that should not be broken after ZWJ if they show up on other systems — as they will inevitably do. I have not verified the Windows changes, but I think we should take it as a serious omen in any event, and try to be more proactive. The UTC ended up making the minimal changes for the current ZWJ sequences. Unfortunately, we didn't push to make a broader change that would protect us against future sequences: aka “future-proofing”. And ZWJ sequences will continue to have the attraction that their fallback behavior is much better than TAG sequences.

The important fact is that for users, the bad effects from breaking too much around ZWJ are way worse than those from breaking too little around ZWJ.

Data

I put together a spreadsheet to illustrate the relevant data: L2/16-094 (Working Draft). The Green cells are cases that work ok after ZWJ, when applying the rules. Here’s a key to the abbreviations:

- GCB, WB, LB are the Grapheme cluster break, Word break, Line break properties.
- GC = General category properties, just fyi
- WNC2 are the Windows characters in the reported Windows sequences:
    - U+1F431+ZWJ+U+1F464/U+1F3CD/U+1F4BB/U+1F409/U+2615/U+1F680

The key rules are:

- http://www.unicode.org/reports/tr29/proposed.html#WB3c
- http://www.unicode.org/reports/tr29/proposed.html#GB11
- http://www.unicode.org/reports/tr14/proposed.html, search for "LB8a Do not break"

Proposal

To future-proof ourselves as much as possible, we'd try to get as many pictographic emoji in the green as we can.

Part 1. The easiest change is to bring Grapheme cluster break and Word break in line with Line break by changing some GCB and WB character properties — no rule changes necessary.

- XX to GAZ for characters in Rows 4-7
    - GAZ and XX already behave identically except around ZWJ.
- EB to EBG for characters in Row 8*
    - EB and EBG already behave identically except around ZWJ.
    - * This change would also allow merging EB and EBG, making the rules/properties actually a bit simpler.

Part 2. The above handles the large majority of characters. For rows 9/10 the GCB and WB change would be just like #1 above; for LB the easiest change is to ID. However, we should not move all of them: some are unsuitable as non-initial character in ZWJ sequence and/or might cause a problem with a line break change. Thus the proposal is to change just the following characters.

37 Characters to be changed to GCB=GAZ, WB=GAZ, and LB=ID

[☠ □ □ □ □ ⚓ ☆ ⚡ ❄ ☄ □ □ □ □ ⌨ ✏ ✒ ⚒ ⚙ ⚗ ⚛ ✂ ⚠ ☢ ☣ ☸ □ ✡ ☯ ☯ † ✝ C☪ ☮ ⚜ □ □]

119 Characters to remain as they are

[♠ ♥ ♦ ♣ ☎ □ ⬭ ⬯ ⬆ ↗ ➡ ↘ ⬇ ↙ ⬅ ↖ ↕ ↩ ↪ □ □ □-□ ♻ □ □ ☑ ✔ ✖ □ □ □-□ □ □ ✳ ✴ ❇ □ □ □-□ □ □ © ® ™ ♈-♓ □ □-□ ▶ □ □ □ ◀ □ □ □ □ □ □ □-□ ⏏ □ □ □ □ □-□ □ □ □-□ ▪ ▫ ■ □ □-□ □ □ □ ○ ● □ □]