L2/22-094R

Regularizing Emoji Structure

2022-04-22 — Mark Davis, Emoji Subcommittee

This is a proposal to regularize the structure for valid emoji, and in so doing allow for additional combinations of valid emoji. Note that being *valid* allows any particular platform, app, or font to support them, but does not make them *RGI*. For the reasons for doing this, and what the implications are, see the Rationale and Background at the end. Here are two examples of what the changes would allow:

Heart + ZWJ + Ukraine-flag

Africa-flag

Importantly, none of these changes are aimed at adding emoji characters or *RGI* emoji. Instead, they are aimed at adding some additional options for *valid* emoji. Many people don't realize the difference between those.

1. The number of ***RGI*** emoji is limited to what is in the current release of emoji. For Unicode 14.0, that is 3633 emoji (including variants). They are generally all supported by the major platforms.
2. The number of ***valid*** emoji is essentially infinite, since they include an indefinitely large number of ZWJ sequences. Few are supported by the major platforms.

Value of Validity

But what is the point, though, if the proposed emoji sequences are not RGI? It is that: *any platform, app, or font can support any valid emoji in plain text.*

Using valid emoji sequences provides for some important features even if an emoji sequence is not generally interchangeable:

- Searching: The emoji sequence can be searched for
- Accessibility: The constituent emoji can be converted to speech
- Fallback: If sent to a recipient that doesn't support it, the sequence of emoji is still a far better fallback than just boxes.
- Parsing: Software can detect that the sequence is intended to be treated as a unit, not just adjacent emoji.

- Evidence for RGI: if a particular valid sequence becomes particularly popular in the future, that factor could be considered.

Proposal

The proposed changes apply to about 10 lines in the UTS #51: Unicode Emoji specification. Below they are indicated with highlighting, with strikethrough for removals.

1. Expand ZWJ emoji with the following changes

The practical impact of these changes is that it allows *any* two or more emoji to be linked together with ZWJ characters, instead of having special exclusions for keycap emoji, flag emoji, and tag sequence emoji (eg, subdivision flags).

This change makes the sequences be valid emoji, but ***does not*** *make them RGI emoji.*

1.4.5 Emoji Sequences

ED-15a. emoji zwj element — An ~~more limited~~ element that can be used in an emoji ZWJ sequence, as follows:

```
emoji_zwj_element :=
~~  emoji_character~~
~~| emoji_presentation_sequence~~
~~| emoji_modifier_sequence~~
| emoji_core_sequence
| emoji_tag_sequence
```

1.4.9 EBNF and Regex

EBNF	Notes
possible_emoji := ~~flag_sequence~~ ~~|~~zwj_element (\x{200D} zwj_element)*	\x{200D} = zero-width joiner
flag_sequence := \p{RI} \p{RI}	\p{RI} = Regional_Indicator
zwj_element := \p{Emoji} emoji_modification? | flag_sequence	

emoji_modification := \p{EMod} | \x{FE0F} \x{20E3}? | tag_modifier	\p{EMod} = Emoji_Modifier \x{FE0F} = emoji VS \x{20E3} = enclosing keycap
tag_modifier := [\x{E0020}-\x{E007E}]+ \x{E007F}	\x{E00xx} are tags \x{E007F} = TERM tag

[Ed Note: the missing tag_modifer is a typo in the original also!]

Regex

```
\p{RI} \p{RI}
| \p{Emoji}
  ( \p{EMod}
  | \x{FE0F} \x{20E3}?
  | [\x{E0020}-\x{E007E}]+ \x{E007F})?
  (\x{200D}
    (\p{RI} \p{RI}
     | \p{Emoji}
       ( \p{EMod}
       | \x{FE0F} \x{20E3}?
       | [\x{E0020}-\x{E007E}]+ \x{E007F})?
    )

  )*
```

2. Fix Emoji Tag sequences with the following changes

The practical impact of this change is that it allows Unicode macroregions (continents and subcontinents), such as *Africa* [002], to be valid emoji flags.

This change makes the sequences be valid emoji, but ***does not*** *make them RGI emoji.*

Annex B: Valid Emoji Flag Sequences

The valid region sequences are specified by 2-letter Unicode region subtags as defined in [CLDR], with idStatus=regular, deprecated, or macroregion. For macroregions, only **UN** and **EU** are valid.

[Ed Note: the above change is not strictly necessary, but helps to avoid confusion.]

C.1 Flag Emoji Tag Sequences

A valid flag emoji tag sequence must satisfy the following constraints:

1. The tag_base and tag_spec are limited to the following:

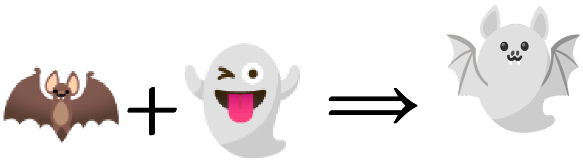
`tag_base`	`U+1F3F4 BLACK FLAG`
`tag_spec`	`(U+E0030 TAG DIGIT ZERO .. U+E0039 TAG DIGIT NINE, U+E0061 TAG LATIN SMALL LETTER A .. U+E007A TAG LATIN SMALL LETTER Z)+`

2. Let SD be the result of mapping each character in the tag_spec to a character in [0-9a-z] by subtracting 0xE0000.
    1. SD must then be a specification as per [CLDR] of either a Unicode subdivision_id or a 3-digit unicode_region_subtag, and
    2. SD must have CLDR idStatus equal to "regular" or "deprecated" or “macroregion”.

Rationale and Background

The regularization of emoji ZWJ sequences both makes the encoded structure of emoji more clear and understandable to developers and users, while it also enables more apps, systems, fonts, etc. to support a wider range of emoji sequences without these sequences needing to be added to the RGI. While keycap sequences are far less likely to be used than flag sequences, there is still benefit in further regularizing the structure to remove exceptional cases.

You can get a peek at how such a valid, non-RGI sequence can function by looking at Emoji Kitchen. There you can take, for example, a bat and a ghost, and produce a combo image:



This, like the other Emoji Kitchen combinations, corresponds to a valid emoji sequence. Although Emoji Kitchen currently emits image-based stickers rather than emoji characters, the high popularity of the Emoji Kitchen use-case of creating novel emojis by combining sequences of existing emojis demonstrates that sequences of valid—yet non-RGI—emoji have a high potential of use.

In addition, stickers like these that have flags could be tagged with a valid emoji ZWJ sequence, such with an HTML img-alt value. That allows for 3 features listed under Value of Validity: searching, accessibility, and fallback (when pasted into plain text).

The purpose of the change to allow macroregions (again, just as valid, not RGI), is to remove a restriction that no longer seems necessary, and would address a demand for certain of them, including Africa.

Note that as always, the Unicode Consortium doesn't require a specific image for emoji character or sequence, even for RGI emoji. The images used in the charts are purely illustrative. That is also true of any newly valid emoji that would result from this change.