

Alternative Encoding Model for Emoji Hair Variations

Author: Charlotte Buff

Mail: irgendeinbenutzername@gmail.com

Submitted: 2017-06-XX

Introduction

After skin colour and gender received systematic presentations in Unicode, hair style is now projected to be the next big step in emoji customization. Since it is not possible to choose an emoji's hair colour or cut in plain text without resorting to invalid variation sequences, unregulated ZWJ hacks, or fragile private-use allocations, many people somewhat understandably feel improperly represented by Unicode emoji as most implementations employ a very plain and generic hair style with fixed colours in their fonts. While these neutral designs should in theory be more than sufficient for interchange, users generally very much prefer having more options rather than less when it comes to emoji.

The UTC tentatively approved four characters representing people with different hair styles – red, white, curly, and bald – for a Unicode 11 release. While public reactions to this decision seemed to be largely positive, I highly suspect that many people do not fully understand how these characters are intended to work in practice and what implications their formal approval would have. It is not unlikely that a significant number of people will read articles with headlines like '*Redhead Emoji Are Finally Coming*' and wrongfully assume that *all* human-form emoji will receive an additional colour option, and then be greatly disappointed and angry when it turns out that that is not remotely the case.

I do not consider the UTC's approach to hair variation to be practical and would like to propose an alternate solution. For easier reading, I will be mostly focussing on red hair as a stand-in for all proposed hair styles in the following discussion.

Flaws in the Current Draft

In the currently agreed-upon model hair style is intrinsic to the character. PERSON WITH RED HAIR is exactly the same as ADULT, just with a more narrow range of acceptable glyphic presentation; while ADULT can have any hair colour – including red –, PERSON WITH RED HAIR must always be ginger. This in my eyes already demonstrates the fundamental flaw with this approach: We are trying to encode glyphic variants of the same abstract character as distinct codepoints.

When the introduction of skin colour variants was initially discussed, one alternative proposal (Everson, L2/14-226) suggested encoding them as atomic characters. This was discarded in favour of the postfix-modifier-based model we use today. Similarly, emoji receive gender not through a set of auxiliary variant characters but by applying the sequences `<200D, 2640, FE0F>` or `<200D, 2642, FE0F>` respectively to the base humans as if they were combining marks (with a few exceptions that were put in place before Emoji 4.0 became a thing). Now to look at hair colour and state that *this* modification has to be implemented with fully separate characters does not strike me as sensible.

ADULT and PERSON WITH RED HAIR are functionally identical; they have the same semantics, but merely slightly different appearances. When searching a document for the character POLICE OFFICER for instance I will find all of its occurrences regardless of which genders and skin tones were chosen by the author. However, searching for ADULT won't find me any people with red, white, curly or no hair as they are all different codepoints with no formal relation between them. This is not good design. Similarly, a user whose device's fonts do not support the sequence "male police officer, Fitzpatrick type 4" will still be able to understand the basic gist of the message because the fallback behaviour shows them a police officer, followed by possibly a missing glyph, followed by the commonly supported character MALE SIGN. The police officer is the important part that carries the most meaning here; skin tone and gender are just bonus information. On the other hand, someone who just received a message containing ten tofus that were all meant to be people with red hair will have no clue what is being said.

Another big issue is the proposed limitation of hair variants to just the generic, featureless people emoji. First and foremost users will not be happy with the lack of children and elderly people with diverse hair. Using the UTC's approach would therefore already necessitate the addition of eight additional characters CHILD WITH RED HAIR, OLDER PERSON WITH RED HAIR and so on for just the basic family members. Many emoji users are children and teenagers, many emoji users are the parents of children and teenagers. Implicitly declaring that children cannot have red hair is not acceptable.

Proposed Implementation

As the previous paragraphs hinted at I'd prefer the UTC to adopt a modifier-based hair model rather than encoding every single hair variant as a separate character. I propose encoding four new characters that act like modifier symbols similar to the already existing Fitzpatrick modifiers.

-  EMOJI HAIR MODIFIER RED COLOUR
-  EMOJI HAIR MODIFIER WHITE COLOUR
-  EMOJI HAIR MODIFIER CURLY
-  EMOJI HAIR MODIFIER BALD

This may initially sound similar to the tag-based approach that was discussed and eventually rejected some time ago, however the crucial difference is that I do not think that Unicode should enable these modifiers for all characters with hair straight away. Instead, the UTC should do what they did with subdivision flags:

Formally validate everything, but only recommend a subset for general interchange until further demand is proven.

Unicode for all intents and purposes now supports thousands of different ISO 3166-2 flags but currently only recommends three of them to vendors. However, all the remaining sequences are still considered valid and can be implemented by anyone at any time without violating the specifications. If these "unofficial" sequences enter widespread use vendors can then request Unicode to simply add the already existing tag sequences to the list of recommendations without requiring any further action or changes to the standard. Hair colour should be dealt with similarly.

At the beginning the list of recommended hair modifier bases should only include the generic family members:

- CHILD, BOY, and GIRL
- ADULT, MAN, and WOMAN
- OLDER ADULT, OLDER MAN, and OLDER WOMAN

This system's main advantage is its flexibility. When – at that is most definitely a ‘when’ and not an ‘if’ – users express considerable demand for other emoji to also have hair variants vendors can readily add the red-haired police officer, the bald dancer, and the person with curls getting a haircut to their fonts and keyboards without having to wait for the next Unicode release. Inversely, if a vendor discovers that supporting five hair styles in addition to six skin colours and three genders is just not feasible within their framework they can simply refuse to support the hair modifiers altogether or only support a minimal subset of sequences while still allowing their users to receive messages from other platforms relatively unscathed, which wouldn't be possible if the hair variants were distinct fundamental characters. Hair variants can be added to and removed from the data files at a whim, something which is definitely not possible with characters.

I understand that this is probably going to lead to an explosion of glyph numbers, the desired avoidance of which was one of the motivations for the termination of TR 52 and its tag system in the first place as far as I am aware, but I am absolutely certain that most users who are wishing for more hair diversity in emoji are not going to be satisfied with the UTC's proposal for very long. Users who aren't aware of the technical background of emoji (which is most of them) will neither understand nor care about the nature of PERSON WITH RED HAIR. People want emoji that they can feel comfortable representing themselves with, and redheads come in more varieties than just “generic adult person”. There will be calls for a ginger police officer, farmer, and athlete in the future and when that time comes the UTC will need to decide between ignoring user demands which would make this whole endeavour pointless in the first place, or designing a hair colour system after all that is bound to be incompatible with past implementations. All of this can be avoided if the modifier characters become the canonical representation of hair right from the beginning.

Some people might be 100% sure in their belief that expanding hair variations to more emoji won't ever be necessary, but things may very well look completely different further down the line. I believe that it is better to create a powerful tool now and potentially “risk” never having to use it than to be forced to come up with some awkward ad-hoc solution in the future if it turns out that the tool indeed was needed. Once PERSON WITH RED HAIR is encoded we cannot get rid of it if we ever decide on a more generalized approach to hair colour afterwards. Maybe I am wrong and the vast majority of users will never want any more red-haired emoji than just Adult, Man, and Woman but the system I propose works for a small selection of characters just as well and remains forwards compatible just in case.

The set of potential hair modifier bases is strictly smaller than the set of all human-form emoji since some characters do not have any visible hair (for example PERSON WITH HEADSCARF), are commonly shown with so little hair that no difference would be visible at standard display sizes either way (for example BABY), or are expected to always be rendered with a very typical hair due (for example FATHER CHRISTMAS). It is unlikely that high demand will arise for some of the lesser used emoji like SLEEPING ACCOMMODATION to receive hair variants, but it is definitely possible. Making hair a dependent modifier means that more variations can be added over time according to user wishes.

Caution should be taken to the fact that hair presence is not consistent across platforms. As an example: Microsoft, Samsung, and HTC show PERSON RAISING BOTH HANDS IN CELEBRATION and PERSON WITH FOLDED HANDS as people with visible hair while most other fonts omit the person and only display a pair of disembodied hands. There is also the interesting edge case of PERSON WITH BLOND HAIR which obviously cannot be shown with red, white or no hair but could theoretically have curls.

Technical Details

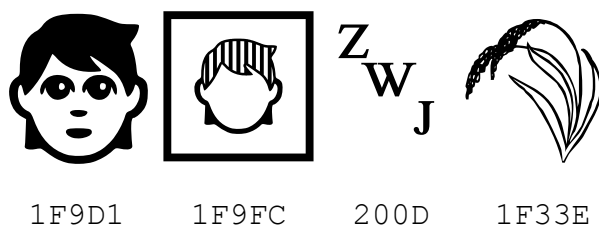
The realisation of this model would require four new characters which for the sake of this proposal I have just allocated to the very end of *Supplemental Symbols and Pictographs*. I highly recommend copying the properties of the Fitzpatrick modifiers, i.e. making hair modifiers spacing, visible symbols. This way recipients can still deduce that some kind of hair modification was intended by the sender even if their device does not support the necessary ligation. Fitzpatrick modifiers were deliberately designed with scenarios like these in mind.

```
1F9FC;EMOJI HAIR MODIFIER RED COLOUR;Sk;0;ON;;;;N;;;;;
1F9FD;EMOJI HAIR MODIFIER WHITE COLOUR;Sk;0;ON;;;;N;;;;;
1F9FE;EMOJI HAIR MODIFIER CURLY;Sk;0;ON;;;;N;;;;;
1F9FF;EMOJI HAIR MODIFIER BALD;Sk;0;ON;;;;N;;;;;
```

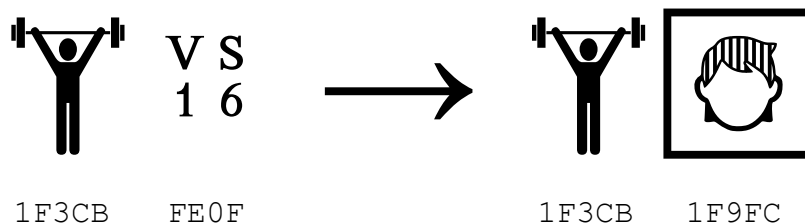
Hair modifiers apply directly to the emoji before them; no ZWJ required. They can be applied to emoji singletons and emoji modifier sequences. That means when both a hair modifier and a Fitzpatrick modifier are present, the skin tone precedes the hair style for backwards compatibility.



Emoji hair modifier sequences can be used in ZWJ sequences. For example, a farmer with red hair:



Hair modifiers automatically imply emoji presentation so no variation selector in case of `Emoji_Presentation=No` is necessary.



A hair modifier sequence should be considered one single grapheme cluster, and lines should not break within them. The easiest solution in my opinion is to simply declare hair modifiers as `Emoji_Modifier=Yes`, and as a result `Grapheme_Cluster_Break=E_Modifier` and `Line_Break=E_Modifier`. Grapheme clustering and line breaking rules would need to be amended to prevent breaking between a pair of `E_Modifiers`. As a side effect any sequence of emoji modifiers would then be treated as a single unit, but since only very specific combinations ever occur in meaningful text this should not cause too many problems. Also this would already future-proof the algorithm to a certain degree should the UTC ever decide to allow more than one Fitzpatrick modifier per character for emoji that show several people. Alternatively new property values specific to hair modifiers could be created to enable more fine-grained control.

It could be useful to introduce a new binary character property `Emoji_Hair_Modifier_Base` to unambiguously mark the set of characters currently recommended to take on hair variants. Of course vendors can always choose to also support additional sequences, just like some fonts already contain Fitzpatrick modifier sequences for characters like HANDSHAKE.

`Emoji_Modifier_Base=No` automatically implies `Emoji_Hair_Modifier_Base=No` but the reverse is obviously not true.