Deprecation Inconsistencies in Code Chart Annotations

Author: Charlotte Buff
Mail: irgendeinbenutzername@gmail.com
Submitted: 2018-09-16

1. Background

As of Unicode 11.0, there are fifteen characters that have the `Deprecated` property, indicating that said characters should be avoided at all costs because their basic concept is fundamentally flawed and/or there exist other characters or character sequences that fulfil their intended purpose much better.

Reading the code charts, however, one could be forgiven for thinking that the set of deprecated characters is substantially larger. Several characters are annotated as not being recommended for any usage without actually having been formally deprecated, which strikes me as a bit of a logical paradox.

I suggest the UTC review the discrepant cases and clarify the situation:

- If the character in question truly should not be used, it ought to be assigned the `Deprecated` property accordingly.
- If the character in question still has valid use cases, its annotations should be refined to clearly state when it is appropriate to use the character and when it should be avoided.

I see no value in this current "two-tier" deprecation system wherein a character can be discouraged but not *really* discouraged, which only leads to confusion about whether it is actually okay to use a certain codepoint or not.

2. List of Characters with Inconsistent Annotations

I identified the following 18 characters that are strongly implied to be deprecated in the code charts, but actually aren't in the UCD. The current, official annotations are provided in italics.

2.1 Combining Diacritical Marks

U+0340: Combining Grave Tone Mark
U+0341: Combining Acute Tone Mark

> ► *»Vietnamese-specific accent placement should be handled instead by specialized rendering of 0300 and 0301. Use of 0340 and 0341 is discouraged.«*
>
> → U+0340 and U+0341 had been deprecated with Unicode 3.2, but were undeprecated in version 5.2 of the standard. This implies that the UTC still sees some value in these two characters and does not want to recommend against their usage entirely. However, this is contradicted by the annotation in the code chart, which clearly states that their canonical equivalents should always be used instead.

U+0344: Combining Greek Dialytika Tonos

> ► *»use of this character is discouraged«*

2.2 Greek and Coptic

U+037E: GREEK QUESTION MARK

► *»003B is the preferred character«*

U+0387: GREEK ANO TELEIA

► *»00B7 is the preferred character«*

2.3 Cyrillic

U+0478: CYRILLIC CAPITAL LETTER UK

► *»for "digraph onik" the preferred spelling is 041E 0443«*

► *»for "monograph uk" the preferred character is A64A«*

U+0479: CYRILLIC SMALL LETTER UK

► *»for "digraph onik" the preferred spelling is 043E 0443«*

► *»for "monograph uk" the preferred character is A64B«*

→ Since there exist other spellings for everything these two characters can represent, there is no reason to ever use them, thus necessitating deprecation.

2.4 Arabic

U+06E1: ARABIC SMALL HIGH DOTLESS HEAD OF KHAH

► *»presentation form of 0652, using font technology to select the variant is preferred«*

2.5 Gujarati

U+0AF1: GUJARATI RUPEE SIGN

► *»preferred spelling is 0AB0 0AC2 0AF0«*

→ U+0AF1 does not decompose into its preferred representation.

2.6 Tibetan

U+0F73: TIBETAN VOWEL SIGN II
U+0F75: TIBETAN VOWEL SIGN UU
U+0F81: TIBETAN VOWEL SIGN REVERSED II

► *»use of this character is discouraged«*

2.7 Khmer

U+17A8: KHMER INDEPENDENT VOWEL QUK

► *»obsolete ligature for the sequence 17A7 1780«*

► *»use of the sequence is now preferred«*

→ U+17A8 does not decompose into its preferred representation.

U+17D3: KHMER SIGN BATHAMASAT

► *»use of this character is discouraged in favor of the complete set of lunar date symbols«*

→ Similar to the Vietnamese tone marks, this character also used to be deprecated until Unicode 5.2, yet the chart still says that it should not be used regardless. If the atomicly encoded lunar date symbols are always preferred, what is this sign meant to be used for instead?

U+17D8: KHMER SIGN BEYYAL

► *»use of this character is discouraged; other abbreviations for et cetera also exist«*

► *»preferred spelling: 17D4 179B 17D4«*

→ U+17D8 does not decompose into its preferred representation.

2.8 Currency Symbols

U+20A4: LIRA SIGN

► *»intended for lira, but not widely used«*

► *»preferred character for lira is 00A3«*

→ If U+00A3 canonically represents lira, what is U+20A4 used for?

2.9 Cyrillic Extended-A

U+2DF5: COMBINING CYRILLIC LETTER ES-TE

► *»preferred representation is the sequence: 2DED 2DEE«*

→ U+2DF5 does not decompose into its preferred representation.

2.10 Sharada

U+111C4: SHARADA OM

► *»use of this character is discouraged«*

► *»recommended sequence is 1118F 11180«*

→ U+111C4 does not decompose into its preferred representation.

3. Unused Diacritical Marks

The following seven characters in the *Combining Diacritical Marks* block could also use improved annotations that better reflect their practical usage:

U+0321: COMBINING PALATALIZED HOOK BELOW

U+0322: COMBINING RETROFLEX HOOK BELOW

U+0334: COMBINING TILDE OVERLAY

U+0335: COMBINING SHORT STROKE OVERLAY

U+0336: COMBINING LONG STROKE OVERLAY

U+0337: COMBINING SHORT SOLIDUS OVERLAY

U+0338: COMBINING LONG SOLIDUS OVERLAY

While superficially appearing like any other combining marks, these characters aren't actually used for producing new letters in practice. No existing characters that contain these diacritics decompose, and new characters with these diacritics are still being encoded atomicly without decomposition mappings as well; the only exception is the usage of U+0338 as a negation stroke for mathematical symbols.

People searching for a way to write (for example) the letter 't' with palatal hook could be misled to believe that the sequence <U+0074, U+0321> is a valid representation of this grapheme, maybe even believing that applying Normalisation Form C will combine the two codepoints into one, when in reality only the stand-alone, normalisation-inert character U+01AB is appropriate for this purpose. Behaviour like this is inconsistent with other attaching diacritics like ogonek or cedilla, which is why it deserves special mention directly in the code charts.

The chart annotations for these combining characters don’t explicitly indicate that something like this may be the case. Worse still, annotations for palatalized hook, retroflex hook, and tilde overlay even describe their meaning in the IPA without any acknowledgement that phonetic transcription actually makes no use of these marks whatsoever, but uses atomic characters instead.

I recommend that comments be added to these characters along the lines of:

This diacritic is not used to form new letters. Use precomposed characters instead.

Additionally, the annotations for U+0338 should indicate that it is still used to negate mathematical symbols despite of this.

I am against formally deprecating these characters entirely, however, since they are still useful for describing the respective diacritic in isolation, for example in explanatory material about transcription systems.