L2/22-266

Title: Summary of SAH and PAG joint meetings
Source: Debbie Anderson (scribe)
Date: 6 October 2022

Report from joint meeting of Script Ad Hoc and Properties and Algorithms Group

Members of the SAH and PAG met twice in September 2022 and discussed a number of issues that intersect the interests of both groups.

Properties in New Proposals

Proposal authors will work with the SAH so their proposals describe key behaviors of the script or characters being proposed (including punctuation), with examples. Such behaviors include directionality, casing, joining behavior (if relevant), segmentation behavior (grapheme cluster/word break/line break); collation; Indic properties (if relevant) and rendering rules in context. Providing as much information as possible will help Unicode experts assign the appropriate properties or verify any suggested property values.

Proposals should provide suggested property values that cover script, general category, bidi class, combining class, case, decomposition mappings, etc. (A forthcoming FAQ in the section “Submitting Successful Character and Script Proposals” discusses this topic.)

To ensure that properties for new proposals are reviewed, the SAH will send a note to PAG when a proposal is deemed “mature” and publicly posted in the document register. At that time, PAG can review the property assignments and see if there are any outstanding issues that need to be addressed. Revisions may be required, depending upon the review.

Modifiers

The addition of modifier letters for phonetic transcriptional systems was an architectural decision made early in the development of Unicode and new modifier additions are being added based on the same criteria.

Modifier letters make sense for a well-established notational system where these things have specific meaning, so that one can put them into databases etc. These modifier letters basically act like diacritics on letters in these systems.

By contrast, complex notational systems such as chemistry or mathematics often use super- and/or subscripting of full expressions, and the use of super- and/or subscripting may also be recursive, with superscripted expressions appearing on elements that are already superscripted, for example. Such cases should be handled with markup and styling, rather than the encoding of individual characters. Abbreviatory conventions have similar considerations, since they may superscript arbitrary strings of text; such conventions should also not be represented with modifier letters.

In sum: Letter modifiers are intended to cover transcriptional phonetic systems (or orthographies derived from them). For more complex notational systems (such as chemical notation and mathematics), styling or markup should be employed instead.

(There will be a forthcoming FAQ in the section on “Ligatures, Digraphs, Presentation Forms vs. Plain Text” that touches on this topic.)

Composites
Composites will still be allowed in cases where an equivalent combining sequence exists

The discussion identified four models of encoding letters with diacritics :
Model 1: only u+¨ (¨ can apply to multiple different base characters)
Model 2: only ü (¨ does not exist on its own, and is not anticipated to apply to multiple chars)
Model 3: both ü and u+¨, canonically equivalent
Model 4: both ü and u+¨, canonically distinct — needs “do not use” tables, not good
By extension, these models are applicable to equivalent combining sequences where the combining marks aren’t diacritics.

All members who voiced their opinion agreed that Model 4 is the least desirable. Model 1, when applied to user-perceived letters, has posed problems for minority communities; for those communities, a simpler model is often better. Also, Peter Constable mentioned that smarter edit controls are still needed, because simple edit controls currently don’t handle whole-sequence backspacing.

In the case of Todhri, the groups agreed with the SAH (section 2 of L2/22-068, based on L2/22-074) and the consensus at UTC [171-C17]: encode EI and U as characters with canonical decompositions (model 3). Note that Todhri, which is not a South Asian script, has acute, inverted breve, and macron as combining mark diacritics -- which can be freely applied to many different base letters (model 1) -- but the dot that appears on EI and U indicates vowel quality and only occurs on those specific base letters.

Future meetings
The groups will again meet if/when issues arise.