L2/19-372

Names for Composition Exclusion Sequences

Submitted by: Asmus Freytag, Michel Suignard, and Mark Davis

Date: October 23, 2019

Certain sequences, particularly in Indic scripts, are canonically equivalent to precomposed letters, but, for reason of composition exclusion represent the normalized forms whether in NFC or NFD.

Much has been made in recent interactions with Unicode of the fact that in some cases the unnormalized forms are the ones typically used.

However, for use cases like documenting the Label Generation Rules (LGR) for IDN domain names, it is necessary to document the NFC forms. These are customarily identified in the documentation both by code point and name. Where sequences are named sequences, the documentation can use the sequence name, otherwise it must fall back on generating something like NAME1 + NAME2 + NAME3….

This is unsatisfactory for composition exclusion sequences, because it hides the fact that they represent a unit of the orthography, and one, furthermore, that is already named (with the name for the precomposed character).

We suggest therefore that Unicode consider adding sequence names for such composition exclusions to the set of named sequences in NamedSequences.txt.

We suggest the following naming pattern:

<script name> SEQUENCE FOR <name of precomposed character minus script name>

For example:

BENGALI SEQUENCE FOR LETTER RRA; 09A1 09BC

BENGALI SEQUENCE FOR LETTER RHA; 09A2 09BC

We propose that the sequences so named cover all decompositions for the entries in Section (1) “Script specifics” in CompositionExclusions.txt. However, it would be acceptable to except composition exclusions for Hebrew on the basis that the names for the precomposed characters themselves are based on components.