L2/19-290

Towards a better documentation of script exemplars

Roozbeh Pournader (WhatsApp)
July 20, 2019

History

This document is written to fulfill UTC Action Item 156-A15 based on discussions during UTC meetings about UTC Document L2/17-430R, as well as discussions with Ken Whistler and Mark Davis outside the meetings.

Background

The two script-related properties, in the UCD, Script and Script_Extensions, are weird, inconsistent, underdefined, and hard to maintain (the author maintains Script_Extensions). There is also no formalized algorithm that uses them, which makes their assigning values more of an art.

There are two major use cases for these properties. One is script itemization, for which various ad hoc algorithms exist in different platforms and is beyond the present discussion.

The second is for font designers and software implementors to figure out which characters they need to cover if they wish to support a certain script. While this is somehow doable using the existing properties, they are far from complete, especially for major scripts like Latin.

And one of the reasons they are far from complete, is the format they are maintained in. ScriptExtentions.txt is not easy to maintain as a source file, since any change to the properties of one character may cause creation of a new group and resorting everything. It’s also hard to figure out the set of characters needed for any script from it without spending a lot of time searching it or writing code that parses it.

Because of this, the author is proposing a new format for maintaining script exemplars. If this format is adopted, Script_Extensions could become a derived property, computed from this and other properties.

Proposal

Adopt a new data file, ScriptExemplars.txt, in the following format:

```
Arabic; 0020..0022 # SPACE..QUOTATION MARK
Arabic; 0025 # PERCENT SIGN
Arabic; 0027..002B # APOSTROPHE..PLUS SIGN
Arabic; 002D..003A # HYPHEN-MINUS..COLON
Arabic; 003C..003E # LESS-THAN SIGN..GREATER-THAN SIGN
Arabic; 005B..005D # LEFT SQUARE BRACKET..RIGHT SQUARE BRACKET
Arabic; 007B # LEFT CURLY BRACKET
Arabic; 007D # RIGHT CURLY BRACKET
Arabic; 00A0 # NO-BREAK SPACE
Arabic; 00AB # LEFT-POINTING DOUBLE ANGLE QUOTATION MARK
Arabic; 00B0..00B1 # DEGREE SIGN..PLUS-MINUS SIGN
Arabic; 00BB # RIGHT-POINTING DOUBLE ANGLE QUOTATION MARK
```

```
Arabic; 00D7 # MULTIPLICATION SIGN
Arabic; 00F7 # DIVISION SIGN
Arabic; 0600..06FF # ARABIC NUMBER SIGN..ARABIC LETTER HEH WITH
      INVERTED V
Arabic; 0750..077F # ARABIC LETTER BEH WITH THREE DOTS HORIZONTALLY
      BELOW..ARABIC LETTER KAF WITH TWO DOTS ABOVE
Arabic; 08A0..08B4 # ARABIC LETTER BEH WITH SMALL V BELOW..ARABIC
      LETTER KAF WITH DOT BELOW
Arabic; 08B6..08C7 # ARABIC LETTER BEH WITH SMALL MEEM ABOVE.. ARABIC
      LETTER LAM WITH SMALL ARABIC LETTER TAH ABOVE
[…]
Arabic; 200C..200D # ZERO WIDTH NON-JOINER..ZERO WIDTH JOINER
Arabic; 2010..2011 # HYPHEN..NON-BREAKING HYPHEN
Arabic; 2018..2019 # LEFT SINGLE QUOTATION MARK..RIGHT SINGLE
      QUOTATION MARK
Arabic; 201C..201D # LEFT DOUBLE QUOTATION MARK..RIGHT DOUBLE
      QUOTATION MARK
Arabic; 2026 # HORIZONTAL ELLIPSIS
Arabic; 2039..203A # SINGLE LEFT-POINTING ANGLE QUOTATION MARK..
      SINGLE RIGHT-POINTING ANGLE QUOTATION MARK
Arabic; 2044 # FRACTION SLASH
Arabic; 204F # REVERSED SEMICOLON
Arabic; 2E41 # REVERSED COMMA
[…]
Arabic; 102E0..102FB # COPTIC EPACT THOUSANDS MARK..COPTIC EPACT
      NUMBER NINE HUNDRED
Arabic; 10E60..10E7E # RUMI DIGIT ONE..RUMI FRACTION TWO THIRDS
[…]
```

Arabic is done as an exercise, to show how vast the exemplar for a major script could be.

Maintenance concerns

As can be noticed from the Arabic example, this is not trivial work. Data can be bootstrapped from the following sources:

- Existing Script and Script_Extensions properties
- UTC Document L2/17-430R
- Existing exemplar data in CLDR
- Mentions in the Unicode Core Spec
- Mentions in Unicode character and script proposals
- Repertoire of system fonts like Noto

The data would need to be expanded and maintained with help from the Unicode community, with the understanding that this is an ongoing project. Future character and script proposers would also be encouraged to provide this information.

If the UTC greenlights the project, the author can prepare a first version of the data file and commit to maintain it for future versions of the Unicode Standard.