Proposal to change the Unicode assigned script values in Scripts.txt for the two Georgian alphabets

Kent Karlsson
2008-11-12

1 ISO 15924 script codes for Georgian scripts

http://www.unicode.org/iso15924/iso15924-codes.html lists two script codes for the two separate scripts that have been/is used for Georgian:

```
Geor     240        Georgian (Mkhedruli)
                    géorgien (mkhédrouli)
Geok     241        Khutsuri (Asomtavruli and Nuskhuri)
                    khoutsouri (assomtavrouli et nouskhouri)
```

Note that these two script codes are mutually exclusive; there is no overlap at all. However, these two scripts are historically related. The characters used for each of these scripts are disjoint. Geor is the modern Georgian caseless script. Geok is the old ecclesiastical cased alphabet.

2 Unicode's Scripts.txt and PropertyValueAliases.txt

Script.txt currently (Unicode 5.1) says w.r.t. Georgian scripts:

```
10A0..10C5    ; Georgian # L&  [38] GEORGIAN CAPITAL LETTER AN..GEORGIAN CAPITAL LETTER HOE
10D0..10FA    ; Georgian # Lo  [43] GEORGIAN LETTER AN..GEORGIAN LETTER AIN
10FC          ; Georgian # Lm       MODIFIER LETTER GEORGIAN NAR
2D00..2D25    ; Georgian # L&  [38] GEORGIAN SMALL LETTER AN..GEORGIAN SMALL LETTER HOE
```

And PropertyValueAliases.txt currently (Unicode 5.1) says, w.r.t. Georgian scripts:

```
sc ; Geor      ; Georgian
```

However, the characters

```
10A0..10C5    ; # L&  [38] GEORGIAN CAPITAL LETTER AN..GEORGIAN CAPITAL LETTER HOE
2D00..2D25    ; # L&  [38] GEORGIAN SMALL LETTER AN..GEORGIAN SMALL LETTER HOE
```

are not of the script with code Geor according to ISO 15924, they are of the script with code Geok. Assigning them to the script code Geor is simply wrong.

3 Corrected script assignments for Scripts.txt

The correct script assignments for the letters of the Georgian scripts are, without using aliases:

```
10A0..10C5    ; Geok # L&  [38] GEORGIAN CAPITAL LETTER AN..GEORGIAN CAPITAL LETTER HOE
2D00..2D25    ; Geok # L&  [38] GEORGIAN SMALL LETTER AN..GEORGIAN SMALL LETTER HOE

10D0..10FA    ; Geor # Lo  [43] GEORGIAN LETTER AN..GEORGIAN LETTER AIN
10FC          ; Geor # Lm       MODIFIER LETTER GEORGIAN NAR
```

With aliases Georgian=Geor and (new) Khutsuri=Geok (but see below):

```
10D0..10FA    ; Georgian # Lo  [43] GEORGIAN LETTER AN..GEORGIAN LETTER AIN
10FC          ; Georgian # Lm       MODIFIER LETTER GEORGIAN NAR

10A0..10C5    ; Khutsuri # L&  [38] GEORGIAN CAPITAL LETTER AN..GEORGIAN CAPITAL LETTER HOE
2D00..2D25    ; Khutsuri # L&  [38] GEORGIAN SMALL LETTER AN..GEORGIAN SMALL LETTER HOE
```

4 Digression: Script code alias naming

UTR 24, Unicode Script property, says:

> Script names are limited to
> - Latin letters A–Z or a–z
> - Digits 0–9
> - SPACE and medial HYPHEN-MINUS
>
> Script names are guaranteed to be unique, even when ignoring case differences and the presence of SPACE or HYPHEN-MINUS. Underscores are not used when assigning script names. Similar restrictions apply to block names.

Yet, in PropertyValueAliases.txt, there are several script code aliases ('script names') that contain underline ('underscore') characters, but none that have space or hyphen-minus in the name. UTR 24 should probably be corrected on this point, rather than changing the script code aliases.

5 New script property value alias(es)

One option would be to keep "Georgian" as alias just for "Geor", and just introduce "Khutsuri" as alias for "Geok". This is close to the ISO 15924 names given for for the Geor and Geok codes. However, it may be desirable to have "Georgian" as alias for "Geor"+"Geok", and introduce aliases "Khutsuri" for "Geok" and "Mkhedruli" for "Geor". Preferably, there should still not be a need to introduce a new code in ISO 15924 for "Geor"+"Geok". If need be, Unicode could use a private-use four-letter script code, such as "Qaag", for "Geor"+"Geok". One could even introduce an additional alias "Mkhedruli_Or_Khutsuri".

An advantage of letting "Georgian" be an alias for "Geor"+"Geok" is that regular expressions like [:script=georgian:] will be stable (matching both Mkhedruli and Khutsuri characters like now), while still allowing [:script=Geor:] and [:script=Geok:] to return proper matches.

So, with this in mind, in Scripts.txt, the relevant assignments should be:

```
10A0..10C5    ; Khutsuri # L&  [38] GEORGIAN CAPITAL LETTER AN..GEORGIAN CAPITAL LETTER HOE
2D00..2D25    ; Khutsuri # L&  [38] GEORGIAN SMALL LETTER AN..GEORGIAN SMALL LETTER HOE
10D0..10FA    ; Mkhedruli # Lo  [43] GEORGIAN LETTER AN..GEORGIAN LETTER AIN
10FC          ; Mkhedruli # Lm       MODIFIER LETTER GEORGIAN NAR
```

And in PropertyValueAliases.txt, the relevant aliases should be:

```
sc ; Geok      ; Khutsuri
sc ; Geor      ; Mkhedruli
sc ; Qaag      ; Mkhedruli_Or_Khutsuri        ; Georgian
```

(That Qaag is Geor+Geok is informal here. Formalising the relation between Qaag and Geor+Geok, or for other similar cases like Hrkt, Jpan, and Kore and what they alias to, is not the topic of this proposal.)

6 Digression: missing entries in PropertyValueAliases.txt

There is an alias entry for Hrkt in PropertyValueAliases.txt:

```
sc ; Hrkt      ; Katakana_Or_Hiragana
```

But the similar cases for Jpan and Kore don't have alias entries, which I think they should have:

```
sc ; Jpan      ; Katakana_Or_Hiragana_Or_Han   ; Japanese
sc ; Kore      ; Hangul_Or_Han                 ; Korean
```

(This still does not formalise the relationships within Unicode.)

7 Effect on IDNA-Update

There is no effect of the suggested change for IETF's IDNA-Update work, since IDNA-Update does not use script codes for table generation or for anything else.

8 Effect on Unicode regular expressions

As mentioned in section 5, if desired "Georgian" can (informally) be made an alias for "Geor"+"Geok" so that [:script=georgian:] can match both Mkhedruli as well as Khutsuri characters like now (they are both arguably Georgian). Using the real script codes, like in [:script=Geor:] and [:script=Geok:], this will and should only match Mkhedruli characters in the first case and only Khutsuri characters in the second case, and should not do erroneous matches (which they do now) given the ISO 15924 scrip code specifications.

9 Effect on OpenType script tags

While OpenType script tags (http://www.microsoft.com/opentype/otspec/scripttags.htm) often appear to look like the ISO 15924 script tag, they are not. Tags such as "bng2" (not four-letter, not in 15924) "jamo" (this does not differ from "Hang" in ISO 15924 and Unicode) and similar cases clearly indicate that they deviate both from ISO 15924 and from Unicode's script assignments to character codepoints. The specification also says that they "correspond to the contiguous character code ranges in Unicode", which clearly does not apply to many scripts (but instead applies to blocks). The latter statement in the OpenType specification may be in error. I still conclude that these script tags are detached both from both ISO 15924 and from Unicode's script assignments for character codes, and as they say: OpenType "Script tags are defined by Microsoft Typography".

10 Stability concerns

The script assignments to character codes are not under a stabilisation policy, and can therefore be changed when appropriate (as in this case). I'm still suggesting to keep the alias "Georgian" in the current meaning. But assigning the script code "Geor" to the Khutsuri characters is in opposition to ISO 15924, and I have the impression that Unicode is striving to follow ISO 15924, and the only deviation would be to use some private use script codes when called for.

The PropertyValueAliases for script codes are also not under a stabilisation policy, and can therefore be changed when appropriate (as in this case).