Partial IP Response to Latin LGR Update

DATE: 2019-12-11

The IP has received and taken note of the Latin LGR Update Report dated Dec 04/05 2019. In the interest of a speedy reply, the IP is limiting its response to significant new material, in particular, there's a formerly missing table now supplied for the first time as part of Section 6.3.4. The IP has issues with or comments on some of the entries and would like to be sure these are available to the GP for consideration.

We intend on commenting on textual/editorial matters more fully at some later point. However, if the GP would like us to have look at some particular sections, please don’t hesitate to point those out.

# General note

In cases where suggested variants appeared doubtful, the IP decided to test them by constructing a possible label and entering it into a variety of address bars (where the user has generally no control over the font selection for each script). This properly evaluates not simply the outline (ink) of a glyph, but its placement and relative size as well, all factors that affect the possible equivalence of simple shapes.

[In general, the Latin GP feels that comparisons should be made using the normal font sizes. Rather than zooming in to make the tiny differences visible.]

[DT: I believe IP’s is making the argument that doing a comparison in the context of a domain name (e.g. the address URL bar, or “label.tld” for that matter) it makes the inspection much more objective because it removes user-determined formatting. For example, in a “latinsecondlevel.arabictoplevel” domain name, the Arabic part will be unfamiliar to a native Latin script user and will certainly look odd when comparing to the Latin part of the domain name e.g. “example. ااا” vs “example.III”.

# Detailed observations

(1) **Cyrillic Palochka**  -- see our discussion of the Cyrillic LGR in the LGR-3 overview

Excerpt from Cyrillic LGR in the LGR-3 overview:

*“The IP reviewed the proposed cross-script variant relation for U+04CF PALOCHKA. The typography for this code point is rather variable across fonts; a homoglyph relation exists most commonly with “l”, (lowercase L), but it can also be rendered as “small-caps i” or as “dot-less i”. Because those code points are not variants of each other, the requirement for transitivity prevents adding additional cross-script variants: the resulting potential for visual confusion is hereby brought to the attention of the String Similarity Review. [Proposal-Cyrillic] lists a number of other code points with varying degrees of visual similarity.”*

(2) **Arabic Letter Alef** – the IP discussed this at length. The tendency seems to be for fonts to show this letter as somewhat floating above the common baseline:

   

 (Firefox on Windows) (Chrome on Mac)

Thus the shape is similar to, but not indistinguishable from Latin small L or dotless i. In the IP’s estimation this would speak against a variant relation, but is perhaps worth calling out as potential confusable.

[This is certainly distinguishable. IF you increase the font size sufficiently. Just as the Caron and Breve are distinguishable at sufficiently large size. But compare abc.lll vs abc.ااا and the difference is far from obvious.]

[DT: IMO. I believe the style difference is noticeable. It can be safely said that there is something not Latin at the right of the dot in abc.ااا]

(3) **Oriya vowel sign Aa** -- as a combining mark it requires the presence of another Oriya letter. Presumably the case is made that alternated with Oriya letter Ttha it mimics Latin ololo.. -- however the Oriya label lacks the alternation in height: ଠାଠାଠ… (We would expect this case to be analyzed in this fashion, i.e. on the basis of the possible conflicting labels, with the conclusion that full variant status is not needed for vowel sign Aa; in contrast, the case for Tha is a bit stronger: an Oriya sequence ଠଠଠ, if presented in any FQDN without another Latin component, could be more easily read as ooo ).

[In the case of Oriya, the most likely conflict would be a series with Letter O and Letter Dotless I. That achieves the same lack of alternating height.]

**latin.latin latin.oriya**

example.oıoıo vs example.ଠାଠାଠ

[DT: The height of the Oriya letters is noticeable higher than that of the Latin ones.]

(4) **Oriya Letter Ttha and Malayalam Letter** Tha -- these appear to take up the full full-height. Should this affect their variant set? Or can an argument be made that this particular shape is "high risk", esp. if presented with a FQDN not containing European x-height letters for comparison (e.g. a Chinese 2nd level domain name combined with .ଠଠଠ as variant for the same with a TLD of .ooo? An explicit argument, e.g. in a footnote, may be required one way or the other.

The IP tends to think that the significant and consistent difference in height argues for these two code points 0D20/0B20 to be in their own variant set, and distinct from Latin o (and its mappings).

[There is the difficulty that, in the minds of users, capital letters and lower case are interchangeable. So Letter O in capitals, producing abc.OOO is, for the users, exactly the same domain as abc.ooo. Which makes the difference in height of Ttha and Tha moot.]

[DT: In the address URL bar the letters are lowercased, so the height difference between latin letters and non-latin letters will be noticeable]

e.g. example.ooo vs. example.ଠଠଠ

(5) **Myanmar Vowel Sign Sgaw Karen** Eu -- the case here is perhaps stronger than for the Oriya vowel sign, as there isn't a height difference to Myanmar letter Wa.

Here is the comparison of the full label: Latin =  ooıoıoı, Myanmar =  (note use of a screenshot for the second one). However, because of the little hook to the left, the Myanmar label might also also close to Armenian = օօւօւօւ.

While there can be a size difference, at smaller font sizes, in browser UIs the only difference remaining is the hook.

 

(firefox on Windows) (Wikipedia search field, Firefox, Windows)

(Left: Latin small o and dotless i)

The IP considers the hook to be a deal-breaker when it comes to making a variant mapping to dotless I, other than via transitivity.
[The hook would be a more convincing factor were it not that the dotless I, in fonts with serifs, comes with a hook as well: ıoıo vs ဝၢဝၢ ]

0101.oioi vs 0101.oıoı vs 0101.ဝၢဝၢ vs oıoı.ဝၢဝၢ

(6) **Hebrew letter vav** -- the hook at the top is rather heavy in any serifed font and would show up well, if the letter is repeated. (If would have to be repeated, if not alternated with the only other candidate Samekh. Note that the latter shows even more distinct signs of a typical Hebrew ductus.) However, we find that some common user interface fonts reduce these characteristics to the point where the code point becomes indistinguishable from lowercase letter dotless I (not L).

  

(Calibri) (Lucida) (FF on windows) (Chrome, Firefox, opera and Safari on Mac)

In light of the choices made by modern sans-serif fonts (used by default in browser user interfaces) identifying vav as a variant of dotless I would seem motivated.

ıoıoıo 

(Latin dotless I and small letter o for comparison) (Latin on the right)

The equivalence isn’t perfect given that in some cases the hebrew letters are shown smaller than the Latin ones, but given that ICANN delegated a .ooo TLD already, the risk seems there.

[Added. Thanks]

Vav vs Dotless I

ו vs ı

(7) **Hebrew letter Samekh** -- like vav, this one appears to have a rather identifiable script ductus in serifed fonts, but lacks any distinction from small letter o in sans-serif.

(See item 6)

ס vs o

(8) **Ethiopic Syllable Pharyngeal A** -- like Hebrew letter Samekh, this one appears to have a rather identifiable script ductus in serifed fonts, and an identifiable shape even in sans-serifed fonts; the IP notes that this was discussed already in the context of LGR-2 and rejected (in the context of a putative variant with Armenian small letter Oh U+0585). We think it should be considered a confusable (at best), and then probably more with digit zero (e.g. on the second level). Worth documenting in an appendix.



(9) **Cyrillic small letter ES** -- the glyph shown is uppercase. Please fix.
(Reviewers noted that there appeared to be issues like that in other parts of the text)

 (10) **Lao vowel sign E** -- like Lao letter Wo, the IP considers the presence of a tell-tale circle at the head of the stroke a strong characteristic of the script ductus for the Lao script. This feature is clearly visible in browser address bars as well. The code point is also a combining mark and can only occur in alternation with Lao Letter Wo. The latter has a different aspect ratio then the putative Latin counterpart, and an even more pronounced leading circle. (We would expect this case to be analyzed in terms of a putative complete label, and, if that is done, it becomes obvious - at least to us - that this is at best a confusable). Latin = ɔcɔc   Lao = ວເ ວເ

[In a normal font size, spotting the little circle isnon-trivial in the mathematical sense:

ɔcɔc  vs ວເວເ

“Obvious”, it isn’t.]

DT: The difference between latin letters and non-latin letters is apparent when compared as a domain name:

example.ɔcɔc vs example.ວເວເ

(11) **Myanmar Letter Nga** -- while this is more circular than Latin C, there are enough Latin fonts where c is fairly circular so that Latin ococ and Mynamar =  (screenshot) look fairly identical

Latin vs Myanmar

c vs င

(12) **Myanmar** **Vowel sign Aa**: Arguably the labels Latin = oɔoɔ and Myanmar =  (screenshot) look nearly identical (there's a slight difference in connectedness, so this proposed variant relation is more marginal than for (11).)

example.oɔoɔ vs example.ဝာဝာ