# Response to PRI 451

submitted by Asmus Freytag, 2022-06-14

After reviewing UTS#39 we found that there are a number of potential omissions in the confusables.txt data file. (For source and background see the end of this document)

Our analysis is based on ICANN's recent publications of Root Zone Label Generation Rules (RZ-LGR) and Second Level Reference LGRs for almost the complete set of "Recommended" scripts.

In these LGRs, a number of characters are considered mutually exclusive with either another character or a character sequence. This determination was made by panels of local experts and users. Where this exclusion is based primarily on appearance, we consider that an omission in the Unicode data file.

We therefore recommend that these be added to the data file before publication.

# Missing Data in Confusables.txt

#### Digit zero

0030	0	0AE6	0	Gujarati digit zero
0030	0	0CE6	0	Kannada digit zero
0030	0	0E50	0	Thai digit zero
0030	0	0ED0	0	Lao digit zero
0030	0	1040	0	Myanmar digit zero
0030	0	17E0	0	? Khmer digit zero
0030	0	0D20	0	Malayalam letter ttha
0030	0	101D	0	Myanmar letter wa – i

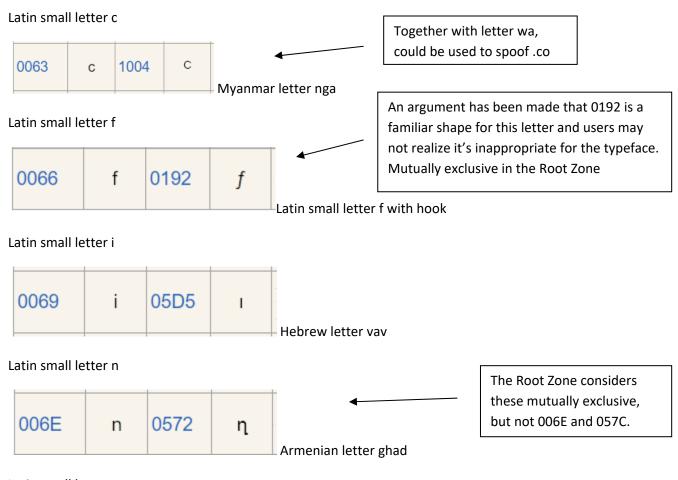
The case can be made that digits are ipso facto confusable semantically, as users may not keep track of which digit set is used in a label when both are available. However, when the shapes are also similar, the potential for confusion increases.

Already in confusables.txt are 09E6 and 0B66 which in the browser used for the screen shots at left look identical to 0AE6 or to 1040.

#### Digit three

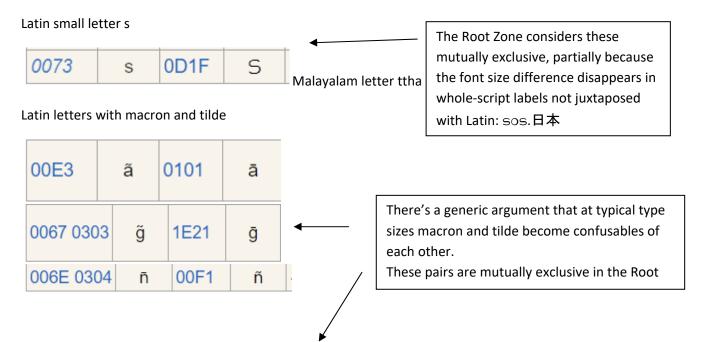
0033	3	0AE9	3	
				Gujarati digit 3

Myanmar letter wa – identical to digit zero



#### Latin small letter p

An argument can be made to consider this confusable with 01BF p Latin small letter wynn. Not only are the shapes close enough but few users know about the wynn and would take it for font idiosyncrasy. Wynn is excluded from the Root Zone for that reason.



00F5	õ	014D	ō
0129	ĩ	012B	ī

# Latin small letter g with dot above

0121 ġ 0123 ģ

Latin small letter g with cedilla

The existing file has 0123 paired with 0127

# Latin small letter eng

014B η 03B7 η 014B η 0572 η

Greek small letter eta

Armenian small letter ghad

# Cyrillic small letter sha

0448 ш 0561 ш

Armenian small letterayb

#### Cyrillic small letter dze

045F µ 1EE5 џ

Latin small letter u with dot below

# Arabic letter alef

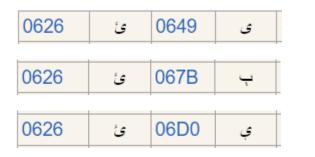
0622	Ī	0623	f .
0622	Ī	0625	ļ.
0622	Ī	0627	1
0622	Ī	0672	į .
0623	Í	0625	\$
0623	Í	0627	1
0623	Í	0672	į .
0625	ļ	0627	1
0625	١	0672	ĺ.
0627	1	0672	į .

The Arabic IDN task force (TFAIDN) concluded that the various forms of Alef should be considered mutually exclusive for domain names.

#### Arabic letter way with hamza above

0624	ؤ	0648	و	Arahic letter way
------	---	------	---	-------------------

#### Arabic letter alef maksura (0649)



TFAIDN concluded that the existing set should be extended to include 0626, 067B and 06D0

#### Arabic letter the marbuta

0629	8	0647	٥
0629	8	06BE	ھ
0629	ъ	06C0	5
0629	8	06C1	٥
0629	8	06C2	هٔ
0629	8	06C3	8
0629	5	06D5	٥

TFAIDN gives the full set as shown here. 06C3 is already listed, but not the other ones. These are mutually exclusive in the Root Zone.

#### Arabic letter teh

Arabic letter tteheh

#### Arabic letter feh

0641	ن	0642	ق	
0641	ف	06A2	į.	

Arabic letter qaf, Arabic letter feh with dot moved below

#### Arabic letter noon



Arabic letter noon ghunna

#### Arabic letter peh

067E	<del>~</del>	06BD	ڻ
067E	<u>_</u>	06D1	ې
067E	<u>_</u>	0752	<u>_</u>

#### Arabic letter nyeh

Arabic letter dyeh

#### Arabic letter dul



#### Devanagari sign candrabindu



#### Devanagari letter a + anusvara

0905 0902	अं	0973	ᆦ	Devanagari letter oe
-----------	----	------	---	----------------------

#### Devanagari digit 2

0968	7	0AB0	5	
0968	२	0AE8	ર	Cuiarati lotto

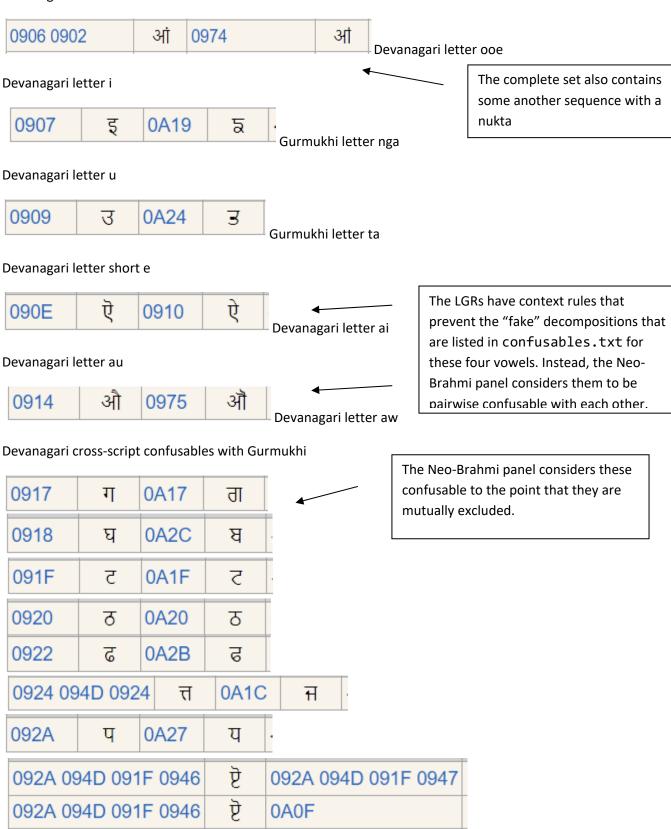
Gujarati letter RA, Gujarati digit 2

#### Devagari nukta

0906	आ	0906 093C	आ्
0913	ओ	0913 093C	ओ
093E	ा	093E 093C	ा
OOOL	OI.	0000 0000	∨!

An argument can be made that the nukta (small dot below) placed on letters where this combination is not expected will not be noticed. The Root Zone makes these mutually exclusive.

#### Devanagari letter aa + anusvara



092D	મ	0A2E	ਮ
092E	म	09AE	ম
092E	म	0A38	ਸ
0935	व	0A15	ਕ
0939	ह	0A35	ਵ
093F	ি	0A3F	ি
0948	ै	0A48	ै
0956	ु	0A41	2
0957	ੂ	0A42	្ជ

Devanagari vowel sign ooe (with and without Anusvara or nukta.

093B	া	093E 0902	ां
093B	া	093E 093C 0902	া

Devanagari cross-script confusables with Bengali



Devanagari in-script and cross-script additional confusables for vowel sign short e

0946	े	0947	े
0946	्र	0A47	े

Devanagari vowel sign au



Devanagari vowel sign au

# Bengali letter ra



# Oriya vowel sign e



# Tamil letter o + lla



# Tamil cross-script confusables with Malaylam

0BAE	Ш	0D25	Ш
0BB5	ഖ	0D16	ഖ
0BC6	െ	0D46	െ
0BC7	ෙ	0D47	ෙ

# Telugu cross-script confuables with Kannada

0C07	ಇ	0C87	ಡ
0C10	80	0C90	න
0C16	ý	0C96	ಖ
0C17	గ	0C97	ヿ
0C1D	ఝ	0C9D	ಝ
0C1F	ట	0C9F	ಟ
0C26	ద	0CA6	ದ
0C28	న	0CA8	ನ

0C30	ŏ	0CB0	ರ
0C33	ళ	0CB3	ಳ
0C3F	ి	0CBF	ి
0C41	ు	0CC1	ు
0C43	ಾ	0CC3	ೃ

Malayalam letter rra: cross-script variants

0D31	n	1002	O
0D31	n	10D8	c

Myanmar letter ga, Georgian letter in

### Sinhala letter iruyanna

0D8D	883	0D9D 0DD8	<del>6</del> 8a
0D8D	883	0DC3 0DD8	æa

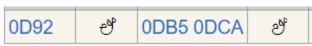
Sinhala combined sequences

#### Sinhala letter eyanna



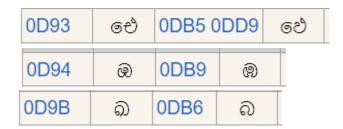
SINHALA LETTER MAHAAPRAANA PAYANNA

# Sinhala letter eeyanna



SINHALA LETTER MAHAAPRAANA PAYANNA + SINHALA SIGN AL-LAKUNA

# Sinhala (additional cases like the prev. three)



0D9D	සි	0DC3	ස
0DA0	ච	0DC0	ව
0DB7	භ	0DC4	හ

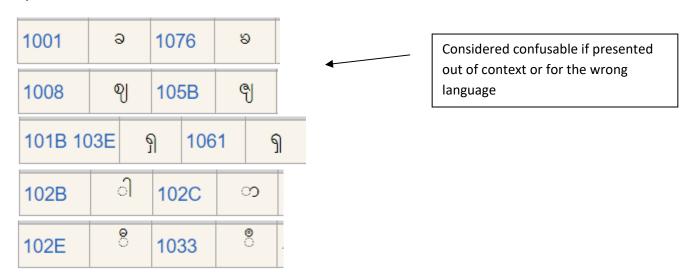
#### Gujarati letter pa



### Myanmar letter k + virama + ka



#### Myanmar variant letter forms



# Myanmar letter nga + asat

1004 103A	ξ	1004 103A 1039	်ံ
1004 103A	δ	105A 103A	ç
1004 103A	င်	105A 103A 1039	ိ

The Root Zone and
Reference LGRs treat
these as mutually
exclusive independent of
how distinctions in
appearance. However,
note that the two
sequences look identical.

Myanmar cross-script confusables with Georgian

1002	n	10D8	С	
1010	တ	10D7	တ	

#### Ethiopic

The Ethiopic script has a number of confusables that are based on phonetic equivalents rather than on visual similarity. The dominant language, Amharic, is commonly spelled phonetically, with apparent free alternation of homophones (for the same word). As if English had a rendom mixture of "lead" / "led", "debt"/"det", or "knight"/"night"/ "nite" and "knite", with all forms equally acceptable in practice. And with the distictions reduced to alternate letters, not sequences.

If this fits the Unicode definition of "confuable", a list can be provided.

Korean Hangul confusables with Han Ideographs

4E2C	k	B258	뉘
723F	爿	B258	뉘
535F	마	B9C8	마
4ECA	今	C2A5	슥
5408	合	C2B4	슴
4E1B	丛	C4F0	쓰
4E15	丕	C870	조
9577	長	D2BD	튽

#### Comments on the sources for this set

These confusables were extracted mainly from Root Zone Label Generations Rules, Version 5 (RZ-LGR 5) a set of script specific repertoires for top-level IDNs that are combined with context rules (that exclude, among others, any sequences Unicode has declared as "do not use"). Any other duplicate spellings (or "close but not quite") have been identified as "variants" and are mutually exclusive. This is equivalent to Unicode's definition of "confusable", except that the focus has been on cases that are either true

substitutions (users without ill intent may substitute one for the other) or those that are considered "practically indistinguishable" on visual grounds.

Some additional confusables were derived from work that ICANN is currently undertaking on Second Level Reference LGRs. These are model LGRs, often extensions of the corresponding Root Zone LGRs, that registries can use on the Second Level, with a similar attention to security.

For both sets, it is assumed that labels are restricted to a single script each, that is, no mixed-script labels are allowed (on the second level, some scripts may have ASCII add ins). However, labels of multiple scripts may coexist on a single zone, so the design includes confusables that can occur between two whole-script labels of different scripts.

There is no need to consider confusables arising from the application of combining marks out of context, because all combining marks are strictly context-limited. (And certainly cannot exist in a mixed script case). Likewise, there is no need to consider cross-script similarity for combining marks, because their allowed base characters are rarely also confusables for the same script pair.

In some cases, the proposal document for a given LGR includes a list of additional confusables that the generation panel of local experts thought did not make the cut. Some of these may well fit the slightly different criteria used in confusables.txt.

In determining their sets of confusables (or the actual visual variants included in the LGRs) the generation panels conducted various levels of research, from informal polling of their own members to formal research by a university. For many of the complex scripts, they also considered extensive lists of conjunct forms. The relevant details are described in their proposal documents for the Root Zone LGRs.

The easiest way to access these, is to go to <a href="https://icann.org/idn">https://icann.org/idn</a> and look for "Root Zone LGR" and there for the list of proposals.

Because the layout of the data differs markedly between LGRs and the confusables.txt, the data for this current report was created by converting the latest confusables.txt into the LGR format and then restricting it to the characters found in the Maximal Starting Repertoire (plus digits and hyphen). The latter represents the de-facto superset of both RZ-LGR 5 and the current set of Second Level Reference LGRs.

This was manually compared to file containing the superset of variant definitions for the aforementioned LGRs, removing duplicates and variants defined for reasons unrelated to visual similarity. (As these are not identified as such, this step could not be automated).