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	૦	Gujarati digit zero
0030	0	0CE6	೦	Kannada digit zero
0030	0	0E50	๐	Thai digit zero
0030	0	0ED0	໐	Lao digit zero
0030	0	1040	၀	Myanmar digit zero
0030	0	17E0	០	? Khmer digit zero
0030	0	0D20	ഠ	Malayalam letter ttha
0030	0	101D	ဝ	Myanmar letter wa – identical to digit zero

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	૩	Gujarati digit 3

Latin small letter c

0063	c	1004	c

Myanmar letter nga

Together with letter wa, could be used to spoof .co

Latin small letter f

0066	f	0192	ƒ

Latin small letter f with hook

An argument has been made that 0192 is a familiar shape for this letter and users may not realize it's inappropriate for the typeface. Mutually exclusive in the Root Zone

Latin small letter i

0069	i	05D5	ו

Hebrew letter vav

Latin small letter n

006E	n	0572	ղ

Armenian letter ghad

The Root Zone considers these mutually exclusive, but not 006E and 057C.

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.

Latin small letter s

0073	s	0D1F	ട

Malayalam letter ttha

The Root Zone considers these mutually exclusive, partially because the font size difference disappears in whole-script labels not juxtaposed with Latin: sos.日本

Latin letters with macron and tilde

00E3	ã	0101	ā
0067 0303	g̃	1E21	ḡ
006E 0304	n̄	00F1	ñ

There's a generic argument that at typical type sizes macron and tilde become confusables of each other.
These pairs are mutually exclusive in the Root

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	ŋ	03B7	η

Greek small letter eta

014B	ŋ	0572	ղ

Armenian small letter ghad

Cyrillic small letter sha

0448	ш	0561	ա

Armenian small letterayb

Cyrillic small letter dze

045F	џ	1EE5	ụ

Latin small letter u with dot below

Arabic letter alef

0622	آ	0623	أ
0622	آ	0625	إ
0622	آ	0627	ا
0622	آ	0672	ٲ
0623	أ	0625	إ
0623	أ	0627	ا
0623	أ	0672	ٲ
0625	إ	0627	ا
0625	إ	0672	ٲ
0627	ا	0672	ٲ

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

Arabic letter wav with hamza above

0624	ؤ	0648	و

Arabic letter wav

Arabic letter alef maksura (0649)

0626	ئ	0649	ى
0626	ئ	067B	ٻ
0626	ئ	06D0	ې

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

Arabic letter the marbuta

0629	ة	0647	ه
0629	ة	06BE	ھ
0629	ة	06C0	ۀ
0629	ة	06C1	ہ
0629	ة	06C2	ۂ
0629	ة	06C3	ۃ
0629	ة	06D5	ە

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

062A	ت	067A	ٺ

Arabic letter tteheh

Arabic letter feh

0641	ف	0642	ق
0641	ف	06A2	ڢ

Arabic letter qaf, Arabic letter feh with dot moved below

Arabic letter noon

0646	ن	06BA	ں

Arabic letter noon ghunna

Arabic letter peh

067E	ﭘ	06BD	ﮊ
067E	ﭘ	06D1	ﯼ
067E	ﭘ	0752	ﭘ

Arabic letter nyeh

0683	ﭽ	0684	ﭾ

Arabic letter dyeh

Arabic letter dul

068E	ڎ	068F	ڏ

Arabic letter dal with three dots above downwards

Devanagari sign candrabindu

0901	ँ	0945 0902	ॅं

Devanagari candra e + anusvara

Devanagari letter a + anusvara

0905 0902	अं	0973	ऽ

Devanagari letter oe

Devanagari digit 2

0968	२	0AB0	૨
0968	२	0AE8	૨

Gujarati letter RA, Gujarati digit 2

Devagari nukta

0906	आ	0906 093C	आ़
0913	ओ	0913 093C	ओ़
093E	ा	093E 093C	ा़
094B	ो	094B 093C	ो़

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

0906 0902	आं	0974	◌ं

Devanagari letter ooe

Devanagari letter i

0907	इ	0A19	ਙ

Gurmukhi letter nga

The complete set also contains some another sequence with a nukta

Devanagari letter u

0909	उ	0A24	ਤ

Gurmukhi letter ta

Devanagari letter short e

090E	ऎ	0910	ऐ

Devanagari letter ai

Devanagari letter au

0914	औ	0975	◌ौ

Devanagari letter aw

The LGRs have context rules that prevent the “fake” decompositions that are listed in `confusables.txt` for these four vowels. Instead, the Neo-Brahmi panel considers them to be pairwise confusable with each other.

Devanagari cross-script confusables with Gurmukhi

0917	ग	0A17	ਗ
0918	घ	0A2C	ਬ
091F	ट	0A1F	ਟ
0920	ठ	0A20	ਠ
0922	ढ	0A2B	ਫ਼
0924 094D 0924	त्त	0A1C	ਜ
092A	प	0A27	ਧ
092A 094D 091F 0946	ऍ	092A 094D 091F 0947	
092A 094D 091F 0946	ऍ	0A0F	

The Neo-Brahmi panel considers these confusable to the point that they are mutually excluded.

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

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

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

Devanagari cross-script confusables with Bengali

093F	ि	09BF	ি

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

0946	ॆ	0947	े
0946	ॆ	0A47	े

Devanagari vowel sign au

094C	ौ	094F	ੌ

Devanagari vowel sign au

Bengali letter ra

09B0	র	09F0	ৰ

Bengali letter ra with middle diagonal

Oriya vowel sign e

0B47	◌େ	1031	◌ေ

Myanmar vowel sign e

Tamil letter o + lla

0B92 0BB3	ஒள	0B94	ஔ

Tamil letter au

Tamil cross-script confusables with Malaylam

0BAE	ம	0D25	ഥ
0BB5	வ	0D16	ഖ
0BC6	◌ெ	0D46	◌െ
0BC7	◌ே	0D47	◌े

Telugu cross-script confuables with Kannada

0C07	ఇ	0C87	ಇ
0C10	ఐ	0C90	ಐ
0C16	ఖ	0C96	ಖ
0C17	గ	0C97	ಗ
0C1D	ఝ	0C9D	ಝ
0C1F	ట	0C9F	ಟ
0C26	ద	0CA6	ದ
0C28	న	0CA8	ನ

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

Malayalam letter rra: cross-script variants

0D31	റ	1002	ဂ
0D31	റ	10D8	ი

Myanmar letter ga, Georgian letter in

Sinhala letter iruyanna

0D8D	ඍ	0D9D 0DD8	ඎ
0D8D	ඍ	0DC3 0DD8	ඎ

Sinhala combined sequences

Sinhala letter eyanna

0D91	එ	0DB5	ඵ

SINHALA LETTER MAHAAPRAANA PAYANNA

Sinhala letter eeyanna

0D92	ඒ	0DB5 0DCA	ඵ්

SINHALA LETTER MAHAAPRAANA PAYANNA + SINHALA SIGN AL-LAKUNA

Sinhala (additional cases like the prev. three)

0D93	ඓ	0DB5 0DD9	ඵෙ
0D94	ඔ	0DB9	ඹ
0D9B	ඛ	0DB6	බ

0D9D	ඝ	0DC3	ඝ
0DA0	ච	0DC0	ච
0DB7	භ	0DC4	භ

Gujarati letter pa

0AAA	પ	0AEB	પ

Gujarati digit five

Compare also to the existing pair 04B7 ч 04CC ч

Myanmar letter k + virama + ka

1000 1039 1000	ကက	1023	ဣ

Myanmar letter i

Myanmar variant letter forms

1001	ခ	1076	ံ
1008	ဈ	105B	ၛ
101B 103E	ဩ	1061	ဩ
102B	ါ	102C	ာ
102E	ိ	1033	ိ

Considered confusable if presented out of context or for the wrong language

Myanmar letter nga + asat

1004 103A	င်	1004 103A 1039	််
1004 103A	င်	105A 103A	ၚ်
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	ဂ	10D8	ი
1010	ဏ	10D7	ნ

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

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 https://icann.org/idn 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).