L2/22-107

Proposal to Add Data for Pairs of Confusable sequences

Asmus Freytag

Proposed:

7 June 2022

This document proposes the addition a number of pairs of confusable sequences to the security (UTS#39) data files in analogy to the “confusables.txt” data collection for single code points.

A separate document proposes a set of eight identical sequences. These same are also absent from the collection of confusables. (Some singletons noted in comments below should also be added).

Background

ICANN recently published the Root Zone Label Generation Rules, Version 5 (see https://icann.org/idn for contents and overview). These Rules, collectively known and RZ-LGR-5, present a dedicated analysis of characters and sequences that users can be expected to freely substitute in the context of domain names. While some of them have strictly identical appearance, a larger set are either not readily distinguished or otherwise seen as alternations that users will substitute. The latter are the source for the current proposal.

The scripts covered in RZ-LGR-5 nearly exhaust the set of “Recommended” scripts from UA#31, while the actual repertoire is limited to characters and sequences that are in common, wide-spread and everyday use. They represent high-priority targets for security exposure but also mitigation efforts. Their absence from the Unicode confusable data represents thus a serious omission that should be remedied.

RZ-LGR-5 represents the culmination of a decade of effort by local volunteer panels composed of native speakers, linguists and technologists; ICANN staff; and a panel of technical experts combining linguistic, Unicode and IDN related expertise. At multiple stages throughout the process, public comments were solicited in a manner similar to the Unicode PRI process.

The rationale for all technical decisions is extensively documented, with each script LGR referencing the underlying proposal document, and, as the case may be, supportive data files.

There are other mappings (singleton to sequence, or singleton to singleton) that also have confusable appearance. This document focuses on the need to document property information, like confusable appearance for cases where it cannot be expressed as a character property, but must truly be considered a property of strings.

None of these mappings have any single code point (in the RZ-LGR-5 repertoire) that would be confusable with the sequence. No data file can be constructed that takes a single code point as the left-hand value. This makes them distinct from the larger set of confusable mappings that can be captured as a property of characters (and of which many, but not all, are already in the security (UTS#39) data files).

The following set of screen captures shows the sequences and their appearance at high resolution

Pairs of Confusable Sequences from RZ-LGR-5

Source	Glyph	Target	Glyph
0906 0901	आँ	0906 093C 0901	आ़ँ
0906 093C 0901	आ़ँ	0911 0902	आँ
0906 0902	आं	0906 093C 0902	आ़ं
092A 094D 091F 0946	ऐ	092A 094D 091F 0947	ऐ
093E 0901	◌ाँ	093E 093C 0901	◌ा़ँ
093E 093C 0901	◌ा़ँ	0949 0902	◌ाँ
093E 0902	◌ां	093E 093C 0902	◌ा़ं

The Neo-Brahmi Generation Panel considers many instances of Nukta confusable: they are not visually prominent and would be ignored by readers not expecting them.

Also possibly confusable with 0974 आं and if so, not a case of pure sequence-to-sequence mapping. However 0974 is not currently listed as confusable

The bare 093E ◌ा is also considered a variant of 093E 093C ◌ा़ and while that is not a pure sequence to sequence mapping, it is not currently listed as confusable and should be added

09A8 09CD 09A5	হ্	09A8 09CD 09B9	ন্হ
09B8 09CD 09A5	হ্	09B8 09CD 09B9	স্হ
0D28 0D4D	ന്	0D7B 0D4D	ൻ്
0D31 0D31	ററ	0D31 0D4D 0D31	റ്റ
0D31 0D31 0D4D 0D31	ററ്റ	0D31 0D4D 0D31 0D31	റ്ററ
0D33 0D33	ള്ള	0D33 0D4D 0D33	ള്ള
0D33 0D33 0D4D 0D33	ള്ളള	0D33 0D4D 0D33 0D33	ള്ളള

For the rationale for this and other similar confusables, please see https://icann.org/idn and look up the Root Zone LGR proposal document for the script.

Also possibly confusable with 0D7B ൻ് and if so, not a case of pure sequence-to-sequence mapping. However 0974 is not currently listed as confusable

The slight kerning caused by the virama was seen as not sufficiently distinct for IDNs.

0D9D 0DD8	ඹa	0DC3 0DD8	ඹa
1004 103A	ငျ	1004 103A 1039	ဋ္ဌ
1004 103A	ငျ	105A 103A	ငြ့
1004 103A	ငျ	105A 103A 1039	ဋ္ဌ
1004 103A 1039	ဋ္ဌ	105A 103A	ငြ့
105A 103A	ငြ့	105A 103A 1039	ဋ္ဌ

Given that 0D9D and 0DC3 are already variants of each other, it's not clear why these sequence needed to be listed. However, neither 0D9D nor 0DC3 are currently listed as confusable, and therefore should be added.

Note that 1004 C and 105A C◌ take on the same appearance if followed by ASAT (see document on identical sequences), and may take on a rather not so distinct appearance if followed by some other diacritic (such as 103D). Shown here bracketed by 101D O on either side for context.

The precise details depend on the renderer making any distinction fragile. The Root Zone LGR considers these two code points variants in any context where they are not followed by a consonant (or digits, if digits were allowed in the Root Zone). confusables.txt does not allow such context-based rules, which would therefore argue for adding all sequences of the form 1004+combining mark and 105A+combining mark, where the range of Myanmar combining marks starts at 102B and ends at 103E, but also the ranges 105E to 1060. 1062 to 1064. and 1081 to 108F.