Proposal to update Intentional.txt

2017-08-17 FOR REVIEW AND ACTION BY THE UTC

Asmus Freytag L2-17/301

The file “Intentional.txt” at http://www.unicode.org/Public/security/10.0.0/intentional.txt contains a list of characters that are considered “identical by intention”, such as Cyrillic “p” and Latin “p”. These characters can be expected to be rendered as identical in a wide range of font styles, which would cause security issues for network identifiers. Intentionally identical characters do exist in the same script as well as across scripts. They also include digits and punctuation.

The definition of “identical by intention” includes the following seven categories of characters that are the result of tradition and encoding decisions, rather than merely an effect of perception of the way characters happen to be rendered. Examples of identical shapes that are not “intentional" in this sense include single stroke glyphs or circles that come out the same across many fonts just because the shapes are so simple.

Examples of character that are identical by intention include the following seven types of characters intentionally disunified based on

1) **script** — such as Latin vs. Cyrillic p and o
2) **directionality** — Arabic vs. Extended Arabic digits
3) **case** — shwa and turned e, D with stroke / cap ETH
4) **digit** or **letter property** — such as shared forms between digits and letters in several scripts
5) **encoding model** artifacts — such as COENG + DA / COENG + TA
6) **non-normalized combining sequences** — many Arabic examples, but also o with stroke, h with stroke
7) **source separation** — mainly in the CJK unified ideographs

Of these, source separation is far from being the least important; it is listed here last, because it is already documented effectively via the Unihan Database. The other six categories are (partially) documented in “intentional.txt”. This document identifies some entries that are missing and proposes them for addition.

There is a wider concern that there are characters that are normally rendered identically in particular fonts or font families commonly used in user interfaces (such as sans-serif), which then creates a potential security risk for identifiers. This document is purposefully limited to a proposal to add certain missing items to the file as it exists today — while recognizing that additional work may be desirable, getting the data collection correct in its current frame of reference is seen as urgent, particularly in light of the work mentioned in L2 17/289.

A secondary recommendation is to improve the documentation of the scope, for example by including the break-down above, and to adopt a better name e.g. “identical by intention” for the collection, and to use that prominently so users don’t have to refer to the file name to know what is meant.

Characters proposed for inclusion into Identical.txt

Each of the following entries fit one of the categories listed above.

Disunified by script

Proposed for addition:

0041; 0410 # (A ~ A) LATIN CAPITAL LETTER A ~ CYRILLIC CAPITAL LETTER A

0042; 0412 # (B ~ B) LATIN CAPITAL LETTER B ~ CYRILLIC CAPITAL LETTER VE

0045; 0415 # (E ~ E) LATIN CAPITAL LETTER E ~ CYRILLIC CAPITAL LETTER IE

0048; 041D # (H ~ H) LATIN CAPITAL LETTER H ~ CYRILLIC CAPITAL LETTER EN

0049; 0406 # (I ~ I) LATIN CAPITAL LETTER I ~ CYRILLIC CAPITAL LETTER BYELORUSSIAN-UKRAINIAN I

004D; 041C # (M ~ M) LATIN CAPITAL LETTER M ~ CYRILLIC CAPITAL LETTER EM

004F ; 041E # (O ~ O) LATIN CAPITAL LETTER O ~ CYRILLIC CAPITAL LETTER O

0050; 0420 # (P ~ P) LATIN CAPITAL LETTER P ~ CYRILLIC CAPITAL LETTER ER

0051 ; 051A # (Q ~ Q) LATIN CAPITAL LETTER Q ~ CYRILLIC CAPITAL LETTER QA[1]

0054; 0422 # (T ~ T) LATIN CAPITAL LETTER T ~ CYRILLIC CAPITAL LETTER TE

0058; 0425 # (X ~ X) LATIN CAPITAL LETTER X ~ CYRILLIC CAPITAL LETTER HA

0057 ; 051C # (W ~ W) LATIN CAPITAL LETTER W ~ CYRILLIC CAPITAL LETTER WE

0059; 04AE # (Y ~ Y) LATIN CAPITAL LETTER Y ~ CYRILLIC CAPITAL LETTER STRAIGHT U

006A; 0458 # (j ~ j) LATIN SMALL LETTER J ~ CYRILLIC SMALL LETTER JE

006F ; 043E # (o ~ o) LATIN SMALL LETTER O ~ CYRILLIC SMALL LETTER O

0071 ; 051B # (q ~ q) LATIN SMALL LETTER Q ~ CYRILLIC SMALL LETTER QA

0077 ; 051D # (w ~ w) LATIN SMALL LETTER W ~ CYRILLIC SMALL LETTER WE

00CF; 0407 # (Ï ~ Ï) LATIN CAPITAL LETTER I WITH DIAERESIS ~ CYRILLIC CAPITAL LETTER YI

[1] Yellow: need further review, may be removed from final proposal. Input welcome.

00D0; 0189 # (Ð ~ Ð) LATIN CAPITAL LETTER ETH ~ LATIN CAPITAL LETTER AFRICAN D

00EF; 0457 # (ï ~ ï) LATIN SMALL LETTER I WITH DIAERESIS ~ CYRILLIC SMALL LETTER YI

0110; 0189 # (Đ ~ Đ) LATIN CAPITAL LETTER D WITH STROKE LATIN CAPITAL LETTER AFRICAN D

0127 ; 045B # (ħ ~ ħ) LATIN SMALL LETTER H WITH STROKE ~ CYRILLIC SMALL TSHE

01B7; 04E0 # (Ʒ ~ Ʒ) LATIN CAPITAL LETTER EZH ~ CYRILLIC CAPITAL LETTER ABKHASIAN DZE

01DD; 04D9 # (ə ~ ə) LATIN SMALL LETTER TURNED E ~ CYRILLIC SMALL LETTER SCHWA

025C; 0437 # (ɜ ~ з) LATIN SMALL LETTER REVERSED OPEN E ~ CYRILLIC SMALL LETTER ZE

0391; 0410 # (Α ~ А) GREEK CAPITAL LETTER ALPHA ~ CYRILLIC CAPITAL LETTER A

0392; 0412 # (Β ~ В) GREEK CAPITAL LETTER BETA ~ CYRILLIC CAPITAL LETTER VE

0395; 0415 # (Ε ~ Е) GREEK CAPITAL LETTER EPSILON ~ CYRILLIC CAPITAL LETTER IE

0397; 041D # (Η ~ Н) GREEK CAPITAL LETTER ETA ~ CYRILLIC CAPITAL LETTER EN

0399; 0406 # (Ι ~ І) GREEK CAPITAL LETTER IOTA ~ CYRILLIC CAPITAL LETTER BYELORUSSIAN-UKRAINIAN I

0399; 04C0 # (Ι ~ Ӏ) GREEK CAPITAL LETTER IOTA ~ CYRILLIC LETTER PALOCHKA

039C; 041C # (Μ ~ М) GREEK CAPITAL LETTER MU ~ CYRILLIC CAPITAL LETTER EM

039F; 041E # (Ο ~ О) GREEK CAPITAL LETTER OMICRON ~ CYRILLIC CAPITAL LETTER O

03A1; 0420 # (Ρ ~ Р) GREEK CAPITAL LETTER RHO ~ CYRILLIC CAPITAL LETTER ER

03A4; 0422 # (Τ ~ Т) GREEK CAPITAL LETTER TAU ~ CYRILLIC CAPITAL LETTER TE

03A5; 04AE # (Υ ~ Ү) GREEK CAPITAL LETTER UPSILON ~ CYRILLIC CAPITAL LETTER STRAIGHT U

03A6; 0424 # (Φ ~ Ф) GREEK CAPITAL LETTER PHI ~ CYRILLIC CAPITAL LETTER EF

03A7; 0425 # (Χ ~ Х) GREEK CAPITAL LETTER CHI ~ CYRILLIC CAPITAL LETTER HA

03B4; 1E9F # (δ ~ ẟ) GREEK SMALL LETTER DELTA ~ LATIN SMALL LETTER DELTA

03B5; 0511 # (ε ~ ԑ) GREEK SMALL LETTER EPSILON ~ CYRILLIC SMALL LETTER REVERSED ZE

03BF; 043E # (ο ~ о) GREEK SMALL LETTER OMICRON ~ CYRILLIC SMALL LETTER O

03F3; 0458 # (ϳ ~ ј) GREEK LETTER YOT ~ CYRILLIC SMALL LETTER JE

0B8E; 0BED # (0B8E ~ 0BED) TAMIL LETTER E ~ TAMIL DIGIT SEVEN

0B95; 0BE7 # (0B95 ~ 0BE7) TAMIL LETTER KA ~ TAMIL DIGIT ONE

0F62; 0F6A # (0F62 ~ 0F6A) TIBETAN LETTER RA ~ TIBETAN LETTER FIXED-FORM RA

1A80; 1A90 # (1A80 ~ 1A90) TAI THAM HORA DIGIT ZERO ~ TAI THAM THAM DIGIT ZERO

Kana

3078; 30D8 # (へ ~ ヘ) HIRAGANA LETTER HE ~ KATAKANA LETTER HE

3079; 30D9 # (べ ~ ベ) HIRAGANA LETTER BE ~ KATAKANA LETTER BE

307A; 30DA # (ぺ ~ ペ) HIRAGANA LETTER PE ~ KATAKANA LETTER PE

Disunified on the base of the Letter property

The following were re-encoded, sometimes as modifier letters, in order to provide characters the letter property. These entries proposed for addition after review by UTC of whether minor rendering alternations in some cases are perhaps systematic:

007C; 01C0 # (| ~ |) VERTICAL LINE ~ LATIN LETTER DENTAL CLICK

01C1; 2016 # (‖ ~ ‖) LATIN LETTER LATERAL CLICK ~ DOUBLE VERTICAL LINE

02B9 ; 2023 # (′ ~ ′) MODIFIER LETTER PRIME ~ PRIME

02BA ; 2033 # (″ ~ ″) MODIFIER LETTER DOUBLE PRIME ~ DOUBLE PRIME

02BB; 2018 # (‘ ~ ‘) MODIFIER LETTER REVERSE APOSTROPHE ~ SINGLE LEFT QUOTATION MARK

02BC; 2019 # (’ ~ ’) MODIFIER LETTER APOSTROPHE ~ SINGLE RIGHT QUOTATION MARK

02BD ; 201B # (‵ ~ ‵) MODIFIER LETTER EVERSED COMMA ~ SINGLE HIGH-REVERSED-9 QUOT…

02C6; 005E # (ˆ ~ ^) MODIFIER LETTER CIRCUMFLEX ACCENT ~ CIRCUMFLEX ACCENT

02C9; 00AF # (¯ ~ ¯) MODIFIER LETTER MACRON ~ MACRON

02CA; 00B4 # (´ ~ ´) MODIFIER LETTER ACUTE ACCENT ~ ACUTE ACCENT

02CB; 0060 # (` ~ `) MODIFIER LETTER GRAVE ACCENT ~ GRAVE ACCENT

02CD; 005F # (_ ~ _) MODIFIER LETTER LOW MACRON ~ LOW LINE

Combining marks

Some combining marks are also arguably disunified. These are proposed for addition:

0306; A67C # (̆ ~ ꙼) COMBINING BREVE ~ COMBINING CYRILLIC KAVYKA

0363; 2DF6 # (ᵃ ~ ⷶ) COMBINING LATIN SMALL LETTER A ~ COMBINING CYRILLIC LETTER A

0364; 2DF7 # (ᵉ ~ ⷷ) COMBINING LATIN SMALL LETTER E ~ COMBINING CYRILLIC LETTER IE

0366; 2DEA # (ᵒ ~ ⷪ) COMBINING LATIN SMALL LETTER O ~ COMBINING CYRILLIC LETTER O

0368; 2DED # (ᶜ ~ ⷭ) COMBINING LATIN SMALL LETTER C ~ COMBINING CYRILLIC LETTER ES

036F; 2DEF # (ˣ ~ ⷯ) COMBINING LATIN SMALL LETTER X ~ COMBINING CYRILLIC LETTER HA

Identical but encoded twice as artifact of encoding model

The Khmer subjoined consonant sequences COENG + DA and COENG + TA render identically, there is really only one symbol. The distinction in code point sequence stems from the encoding model (using an operator sequence of COENG + CONS for subjoined CONS, instead of enumerating the latter). Proposed for addition.

17D2 178A; 17D2 178F # (◌្ដ ~ ◌្ត) KHMER COENG + DA ~ KHMER COENG + TA

Arabic Digits

The file as it stands contains only cases where digits are intentionally identical to letters. However, these cases from the two sets of Arabic digits are not really in a different class of intentionality, and digits are usually part of identifiers. These should be added.

0660; 06F0 # (٠ ~ ۰) ARABIC-INDIC DIGIT ZERO ~ EXTENDED ARABIC-INDIC DIGIT ZERO

0661; 06F1 # (١ ~ ۱) ARABIC-INDIC DIGIT ONE ~ EXTENDED ARABIC-INDIC DIGIT ONE

0662; 06F2 # (٢ ~ ۲) ARABIC-INDIC DIGIT TWO ~ EXTENDED ARABIC-INDIC DIGIT TWO

0663; 06F3 # (٣ ~ ۳) ARABIC-INDIC DIGIT THREE ~ EXTENDED ARABIC-INDIC DIGIT THREE

0667; 06F7 # (٧ ~ ۷) ARABIC-INDIC DIGIT SEVEN ~ EXTENDED ARABIC-INDIC DIGIT SEVEN

0668; 06F8 # (٨ ~ ۸) ARABIC-INDIC DIGIT EIGHT ~ EXTENDED ARABIC-INDIC DIGIT EIGHT

0669; 06F9 # (٩ ~ ۹) ARABIC-INDIC DIGIT NINE ~ EXTENDED ARABIC-INDIC DIGIT NINE

Existing Entry: PALOCHKA

There is a significant issue with one of the entries

026A ; 04CF # (ɪ ~ I) LATIN LETTER SMALL CAPITAL I ~ CYRILLIC SMALL LETTER PALOCHKA

While the existing entry above may have reflected some original intention, the actual situation seems to be that many fonts, and not exclusively sans-serif fonts use the same (or nearly the same) glyph for these two characters instead:

006C; 04CF # (l ~ I) LATIN LETTER SMALL L ~ CYRILLIC SMALL LETTER PALOCHKA

Overall, the relation appears to be best represented by mapping to capital letters:

0049; 04C0 # (I ~ I) LATIN CAPITAL LETTER I ~ CYRILLIC LETTER PALOCHKA

0406; 04C0 # (І ~ I) CYRILLIC CAPITAL LETTER BYELORUSSIAN-UKRAINIAN I CYRILLIC LETTER PALOCHKA

The recommendation would be to remove the existing mapping and instead to add the two mappings to capital letters. The mapping to lowercase L would be appropriate for a future file tracking a different property such as “not reliably distinct”.

Existing Entries: Inverted Column Order

The following existing entries are listed in the latest file in reverse column order: all other entries have the smaller code point in the first column and the larger in the second column. This should be corrected.

1D0D ; 043C # (м ~ М) LATIN LETTER SMALL CAPITAL M ~ CYRILLIC SMALL LETTER EM

1D1B ; 0442 # (т ~ Т) LATIN LETTER SMALL CAPITAL T ~ CYRILLIC SMALL LETTER TE

2C67 ; 04A2 # (Ⱨ ~ Ң) LATIN CAPITAL LETTER H WITH DESCENDER ~ CYRILLIC CAPITAL LETTER EN WITH DESCENDER

2C69 ; 049A # (Ⱪ ~ Қ) LATIN CAPITAL LETTER K WITH DESCENDER ~ CYRILLIC CAPITAL LETTER KA WITH DESCENDER

A9D0 ; A9C6 # (꧐ ~ ꧆) JAVANESE DIGIT ZERO ~ JAVANESE PADA WINDU

Non-normalizable combining sequences

Many common letters, such as O WITH STROKE, H WITH STROKE etc. intentionally lack a decomposition. It is possible to achieve more or less closely matching appearance – sometimes identical, by using some of the combining hook below or stroke, tilde or solidus overlay characters.

There are two alternatives for documenting this intent.

1) Document the combining marks that will not be used in decompositions
2) Document the “pseudo” decompositions for these letters

This proposal recommends alternative 1 (documenting the affected combining marks) for these reasons:

- the number of entries is smaller and can simultaneously serve to document encoding policy
- the preferred way to mitigate security risks is to exclude combining marks that are not required for identifiers (even if they are formally PVALID in the protocol).

- the list does not need to updated for every addition of a “precomposed” character

The characters proposed for this documentation include

0321 # COMBINING PALATALIZED HOOK BELOW

0322 # COMBINING RETROFLEX HOOK BELOW

0334 # COMBINING TILDE OVERLAY

0335 # COMBINING SHORT STROKE OVERLAY

0336 # COMBINING LONG STROKE OVERLAY

0337 # COMBINING SHORT SOLIDUS OVERLAY

0338 # COMBINING LONG SOLIDUS OVERLAY

Some, perhaps even ALL combining marks for the Arabic script are similarly affected by a policy of non-normalizable combining sequences. The proposal is to document ALL of them, but with a recommendation to the UTC to review this against the current state of the encoding practice for Arabic.

Characters that are not reliably distinct

The following collections of related characters are presented here because they appear **not** to fit the criteria for inclusion into “intentional.txt”. However, they have value as counterexamples or as a seed for some future project focused on separately identifying cases where some significant subsets of fonts (important for user interfaces) use an identical glyph; this status could be expressed as “not reliably distinct”.

Cross-Script letters that often show some variation in typography

0278; 0444 # (ɸ ~ ф) LATIN SMALL LETTER PHI ~ CYRILLIE SMALL LETTER EF

03C6 ; 0444 # (ɸ ~ ф) GREEK SMALL LETTER PHI ~ CYRILLIC SMALL LETTER EF

While presentation using the font here will make all four glyphs above look identical, they are more commonly rendered with some distinction. (Examples 1 and 2 show the two common alternate shapes for GREEK SMALL LETTER PHI, while 3 and 4 are Latin and Cyrillic).

1. αφία
2. αφία
3. αφία
4. αφία

As a result, these entries are not proposed for “identical.txt” but might make the cut for some other status that is more focused on the de-facto rendering across UI fonts, such as “not reliably distinct”.

03C6; 3D5D # (φ ~ ϕ) GREEK SMALL LETTER PHI ~ GREEK PHI SYMBOL

Of the two characters above, U+03D5 is not PVALID in IDNA 2008; it is listed here for comparison, because fonts will alternate the glyph assignment between U+03C6 and U+03D5; both forms are recognized by Greek users as the same letter; either one is read as PHI. This is one of several examples where two characters were disunified because mathematical and technical use had given the two glyph forms alternate interpretation, but in general text use, they are as fully interchangeable as “bowl a” and “hook a” in Latin. Some fonts either switch the glyph assignment between the character codes or show both of them the same way. (The example as shown follows the recommendation in the code charts, but note that this is not true in the earlier case for 03C6).

Because both forms of PHI are taken as identical by the native reader, any other character that looks identical to ϕ can be substituted in an identifier even if the font normally uses the other form for Greek. This is an example where it is not enough to simply look at the code tables for glyphs of PVALID code points in order to spot security risks based on glyphs looking identical.

Armenian

The chart font used for Armenian is in a font style that is particular to the Armenian script, using very distinct shapes for the characters. However, actual system fonts, whether serif or sans-serif, are using styles that harmonize with other European typography.

In these type styles, certain letters either use identical glyphs, or glyphs that are not recognizable as Armenian.

004F ; 0555 # (O ~ Ο) LATIN CAPITAL LETTER O ~ ARMENIAN CAPITAL LETTER OH

0053 ; 054F # (S ~ Տ) LATIN CAPITAL LETTER S ~ ARMENIAN CAPITAL LETTER TIWN

0055 ; 054D # (U ~ Ս) LATIN CAPITAL LETTER U ~ ARMENIAN CAPITAL LETTER VO

006F ; 0585 # (o ~ օ) LATIN SMALL LETTER O ~ ARMENIAN SMALL LETTER OH

043E; 0585 # (о ~ օ) CYRILLIC SMALL LETTER O ~ ARMENIAN SMALL LETTER OH

In particular, the glyphs for these first four of these differ minimally, even in the traditionalist style used in the code charts. At a minimum, these cases would qualify for a status of “not reliably distinct”, because, while they are distinct in the traditionalist font style, they are not distinct in common system and UI fonts. The decision of whether to include these in “intentional.txt” would hinge on whether O and OH, but also U and VO are seen as having a deep relation beyond surface appearance. The recommendation to the UTC is to review this aspect.

The following are typically rendered with some slight distinction even in modern system fonts.

0067 ; 0581 # (g ~ ɡ) LATIN SMALL LETTER G ~ ARMENIAN SMALL LETTER CO

0068 ; 0570 # (h ~ հ) LATIN SMALL LETTER H ~ ARMENIAN SMALL LETTER HO

006E ; 0578 # (n ~ ո) LATIN SMALL LETTER N ~ ARMENIAN SMALL LETTER VO

0071 ; 0566 # (q ~ զ) LATIN SMALL LETTER Q ~ ARMENIAN SMALL LETTER QA

0075 ; 057D # (u ~ ս) LATIN SMALL LETTER U ~ ARMENIAN SMALL LETTER SEH

In the font used here, the Armenian characters tend to show serifs, while the Latin characters do not; in the context of a serifed font for Latin, this distinction would disappear. For SMALL G and Q, these distinctions are more noticeable and might lead one to classify them as merely confusable. From a security perspective, SMALL H, N and U would qualify as “not reliably distinct”, because the distinction would be lost for serifed fonts. However, because there seems to always be a residual difference, however small, it is not possible to consider these as identical. Not proposed, however UTC may recommend them for review as part of a future project to capture “not reliably distinct” characters.