Confusing Character Shapes for Arabic Script Block

Arabic Script IDN Working Group (ASIWG)
Version 1.02
26 May 2010

Revision History:

Date	Version	Changes	Changed by
28 Aug 2008	1.0	Created	Arfan Mansoor
02 Sep 2008	1.01	Review and Formatting	Arfan Mansoor and Atif Gulzar
26 May 2010	1.02	Reviewed	Rabia Sirhindi and Sarmad Hussain

1. Objective

Arabic script is cursive in nature and each letter can have four different shapes: initial, medial, final and isolated according to its context in ligature. There are many characters in Arabic block of Unicode (U+06XX and U+0750 to U+077F) that are similar in one or more of these forms and these visual similarities between characters can cause confusions and may be major source of spoofing. The purpose of this document is to identify these confusing character shapes at different positions of ligature.

2. Categories

These confusing characters are divided into three different categories:

i) Characters with same shape not distinct in any language

ii) Confusable characters with similar shapes for at least one language but distinct in at least one language

iii) Confusable characters with similar shapes not distinct in any language

2.1 Characters with same shapes not distinct in any language

These characters are similar in shape at one or more positions and therefore they are confusable for any language using these characters. For example U+0649 (ى) and U+06CC (ی) have same isolated forms, whereas U+0643 (ك) and U+06A9 (ک) have same initial and medial forms although they have different isolated forms. All of these characters are discussed below.

2.2.1 'ک' Based Characters

ARABIC LETTER KAF (U+0643) and ARABIC LETTER KEHEH (U+06A9) which is used in Persian and Urdu are confusing at their initial and medial positions.

Table 1: 'ک' Based Characters

Unicode	Initial Form	Medial Form	Final Form	Isolated Form
U+06A9 (ک)	ک‍	‍ک‍	‍ک	ک

U+0643 (ك)	ﻛ	ﻜ	ﻚ	ك

2.2.2 ‘ه’ Based Characters

HEH based characters are very confusing in Arabic script. ARABIC LETTER HEH (U+0647) and ARABIC LETTER HEH DOACHASHMEE (U+06BE) for Urdu are confusing at their initial, medial and isolated positions, while ARABIC LETTER HEH (U+0647), ARABIC LETTER HEH GOAL (U+06C1) for Urdu, and ARABIC LETTER AE (U+06D5) are confusing at their isolated positions.

Table 2: ‘ه’ Based Characters

Unicode	Initial Form	Medial Form	Final Form	Isolated Form
U+0647 (ه)	ﻫ	ﻬ	ﻪ	ه
U+06BE (ھ)	ﮬ	ﮭ	ﮭ	ﮬ
U+06C1 (ہ)	ﮨ	ﮩ	ﮧ	ہ
U+06D5 (ە)	-	-	ـە	ە

2.2.3 ‘ي’ Based Characters

Although ARABIC LETTER FARSI YEH (U+06CC) does not have dots in isolated form, but it has two dots below at initial and medial position and confuse with ARABIC LETTER YEH (U+064A) at these positions. Furthermore, its isolated form is confusing with ARABIC LETTER ALEF MAKSURA (U+0649).

Table 3: ‘ي’ Based Characters

Unicode	Initial Form	Medial Form	Final Form	Isolated Form
U+064A (ي)	ﻳ	ﻴ	ﻲ	ي
U+06CC (ی)	ﯾ	ﯿ	ﯽ	ی

U+0649 (ى)	-	-	ـى	ى

2.2.4 'ڧ' Based Characters[1]

ARABIC LETTER QAF WITH DOT ABOVE (U+06A7) and ARABIC LETTER FARSI FEH (U+0641) confuse at all (initial, medial, final, isolated) positions.

Table 4: 'ڧ' Based Characters

Unicode	Initial Form	Medial Form	Final Form	Isolated Form
U+06A7 (ڧ)	ڧـ	ـڧـ	ـڧ	ڧ
U+0641 (ف)	فـ	ـفـ	ـف	ف

2.2.5 'ة' Based Characters

ARABIC LETTER TEH MARBUTA (U+0629) and ARABIC LETTER TEH MARBUTA GOAL (U+06C3) are confusing at their isolated positions.

Table 5: 'ة' Based Characters

Unicode	Initial Form	Medial Form	Final Form	Isolated Form
U+0629 (ة)	-	-	ـة	ة
U+06C3 (ۃ)	-	-	ـۃ	ۃ

2.2.6 'ۀ' Based Characters

ARABIC LETTER HEH WITH YEH ABOVE (U+06C0) and ARABIC LETTER HEH GOAL WITH HAMZA ABOVE (U+06C2) are confusing at their isolated positions.

[1] In Arial Unicode MS, character under Numbering 2.2.4, 2.2.7, and 2.2.8 do not combine at their medial positions, but in other fonts like Tahoma (as in our examples) and Times New Roman they combine at their medial positions and create confusions.

Table 6:'ۀ' Based Characters

Unicode	Initial Form	Medial Form	Final Form	Isolated Form
U+06C0 (ۀ)	-	-	ـۀ	ۀ
U+06C2 (ۂ)	-	-	ـۂ	ۂ

2.2.7 'ٹ' Based Character

ARABIC LETTER RNOON (U+06BB) and ARABIC LETTER TTEH for Urdu (U+0679) are confusing at their initial, final, and isolated positions.

Table 7: 'ٹ' Based Character

Unicode	Initial Form	Medial Form	Final Form	Isolated Form
U+06BB (ڻ)	ڻـ	ـڻـ	ـڻ	ڻ
U+0679 (ٹ)	ٹـ	ـٹـ	ـٹ	ٹ

2.2.8 'ث' Based Characters

ARABIC LETTER NOON WITH THREE DOTS ABOVE (U+06BD) and ARABIC LETTER THEH (U+062B) are confusing at their all (initial, medial, final, isolated) positions.

Table 8: 'ث' Based Characters

Unicode	Initial Form	Medial Form	Final Form	Isolated Form
U+06BD (ڽ)	ڽـ	ـڽـ	ـڽ	ڽ
U+062B (ث)	ثـ	ـثـ	ـث	ث

2.2.9 Confusing digits

Unicode provides two sets of digits for Arabic script. One from U+0660 to U+0669 is for Arabic digits and second from U+06F0 to U+06F9 is used for Eastern Arabic-Indic digits for languages of Iran, Pakistan and India (Persian, Sindhi, Urdu etc.). Confusing character with same shapes from these two blocks are listed below.

Table 9: Confusing digits (Arabic-Indic and Eastern Arabic-Indic)

Arabic-Indic	Eastern Arabic-Indic
٠(U+0660)	۰(U+06F0)
١ (U+0661)	۱ (U+06F1)
٢ (U+0662)	۲ (U+06F2)
٣ (U+0663)	۳ (U+06F3)
٥ (U+0665)	۵ (U+06F5)
٧ (U+0667)	۷ (U+06F7)
٨ (U+0668)	۸ (U+06F8)
٩ (U+0669)	۹ (U+06F9)

In case of digits we have one more problem. The codepoints (06F0 to 06F9) are defined for Persian, Sindhi and Urdu, but some of these characters have different shapes e.g. U+06F4, U+06F6, and U+06F7 have different glyphs in Persian than Sindhi and Urdu while U+06F4 has different glyphs in Sindhi, and Urdu.

Table 10: Confusing digits (Eastern Arabic-Indic)

Urdu and Sindhi	Persian
۰(U+06F0)	۰(U+06F0)
۱ (U+06F1)	۱ (U+06F1)
۲۲ (U+06F2)	۲ (U+06F2)
۳ (U+06F3)	۳ (U+06F3)
۴ (U+06F4) (Urdu)	۴ (U+06F4)
۴ (U+06F4) (Sindhi)	

۵ (U+06F5)	۵ (U+06F5)
۶ (U+06F6)	۶ (U+06F6)
۷ (U+06F7)	۷ (U+06F7)
۸ (U+06F8)	۸ (U+06F8)
۹ (U+06F9)	۹ (U+06F9)

2.2 Confusable characters with similar shapes for at least one language but distinct in at least one language

There are characters that distinct in particular language but they may be confusable for users of other languages. For example U+06A9 (ک) and U+06AA (ڪ) are distinct Sindhi letters but confusing for Urdu speakers. Similarly U+064A (ي) and U+06CC (ی) are distinct Pashto letter but are confusing for Urdu speakers. These letters are discussed below:

2.2.1 ‘ک’ Based Character

ARABIC LETTER KEHEH (U+06A9) and ARABIC LETTER SWASH KAF (U+06AA) and ARABIC LETTER KAF (U+0643) are shape variants in Arabic. First two characters are confusing in Urdu but in Sindhi these are distinct characters.

Table 11: ‘ک’ Based Character

Unicode	Initial Form	Medial Form	Final Form	Isolated Form
U+06A9 (ک)	کـ	ـکـ	ـک	ک
U+06AA (ڪ)	ڪـ	ـڪـ	ـڪ	ڪ
U+0643 (ك)	كـ	ـكـ	ـك	ك

2.2.2 ‘ی’ Based Character

ARABIC LETTER FRASI YEH (U+06CC) for Persian and Urdu is confusing with ARABIC LETTER YEH (U+064A) at its initial and medial positions and with ARABIC LETTER ALEF MAKSURA (U+0649) at its final and isolated positions.

Table 12: 'ی' Based Character

Unicode	Initial Form	Medial Form	Final Form	Isolated Form
U+064A (ي)	يـ	ـيـ	ـي	ي
U+06CC (ی)	یـ	ـیـ	ـی	ی
U+0649 (ى)	-	-	ـى	ى

2.2.3 ALEF with HAMZA ABOVE letters

ARABIC LETTER ALEF WITH HAMZA ABOVE (U+0623) and ARABIC LETTER ALEF WITH WAVY HAMZA ABOVE (U+0672) for Baluchi and Kashmiri are confusing in their final and isolated forms.

Table 13: ALEF with HAMZA ABOVE letters

Unicode	Initial Form	Medial Form	Final Form	Isolated Form
U+0623 (أ)	-	-	ـأ	أ
U+0672 (ٲ)	-	-	ـٲ	ٲ

2.2.4 ALEF with HAMZA BELOW letters

ARABIC LETTER 'ALEF WITH HAMZA BELOW' (U+0625) and ARABIC LETTER 'ALEF WITH WAVY HAMZA BELOW' for Baluchi and Kashmiri (U+0673) are confusing at their final and isolated positions.

Table 14: ALEF with HAMZA BELOW letters

Unicode	Initial Form	Medial Form	Final Form	Isolated Form
U+0625 (إ)	-	-	ـإ	إ
U+0673 (ٳ)	-	-	ـٳ	ٳ

2.2.5 Shape variant confusion

In addition to these categories there may be confusable characters within some language. For example U+06D2 (ے) and U+064A (ي) are confusable for Arabic speakers even though these are not similar in shape, in Arabic these are considered stylistic variant of same character U+064A.

2.3 Confusable characters with similar shapes not distinct in any language

There are characters in different languages with slight variations e.g. vertical dots vs. horizontal dots, three dots pointing downwards vs. three dots pointing upward below etc. These characters can also create confusions for users. These characters are listed below.

Table 15: Confusing Characters with Similar Shapes

Unicode	Characters	Remarks
i) U+062A ii) U+067A	i) ت ii) ٺ	i) ARABIC LETTER THE ii) ARABIC LETTER TTEHEEH
i) U+062B ii) U+067D	i) ث ii) ٽ	i) ARABIC LETTER THEH ii) ARABIC LETTER THE WITH THREE DOTS (Sindhi)
i) U+063C ii) U+0764	i) ؼ ii) ݤ	i) ARABIC LETTER KEHEH WITH THREE DOTS BELOW ii) ARABIC LETTER

		KEHEH WITH THREE DOTS BELOW POINTING UPWARDS BELOW
i) U+064A ii) U+06D0	i) ي ii) ې	i) ARABIC LETTER YEH ii) ARABIC LETTER E (Pashto and used as letter bbeh in Sindhi)
i) U+067E ii) U+0752	i) پ i) ݒ	i) ARABIC LETTER PEH (Urdu, Persian) ii) ARABIC LETTER BEH WITH THREE DOTS PONITNG UPWARDS BELOW
i) U+0683 ii) U+0684	i) ڃ ii) ڄ	i) ARABIC LETTER NYEH (Sindhi) ii) ARABIC LETTER DYEH (Sindhi)
i) U+0686 ii) U+0758	ii) چ iii) ݘ	i) ARABIC LETTER TCHEH (Urdu, Persian) ii) ARABIC LETTER HAH WITH THREE DOTS POINTING UPWARDS BELOW
i) U+068E ii) U+068F	i) ڎ ii) ڏ	i) ARABIC LETTER DUL (Burushaski) ii) ARABIC LETTER DAL WITH THREE DOTS ABOVE DOWNWORDS (Sindhi)
i) U+06A0 ii) U+075F	i) ڠ iv) ݟ	i) ARABIC LETTER AIN WITH THREE DOTS ABOVE ii) ARABIC LETTER AIN WITH THREE POINTING DOWNWORDS ABOVE
i) U+06B2 ii) U+06B3	i) ڲ ii) ڳ	i) ARABIC LETTER GAF WITH TWO DOTS BELOW ii) ARABIC LETTER GUEH (Sindhi)

i) U+075D ii) U+075E	i) ݝ ii) ݞ	i) ARABIC LETTER AIN WITH TWO DOTS ABOVE ii) ARABIC LETTER AIN WITH TWO DOTS VERTICALLY ABOVE
i) U+0697 ii) U+076B	i) ڗ ii) ݫ	i) ARABIC LETTER REH WITH TWO DOTS ABOVE (Dragwa) ii) ARABIC LETTER REH WITH TWO DOTS VERTICALLY ABOVE (Torwali, Ormuri)
i) U+0649 ii) U+06CD iii) U+06CC	i) ى ii) ۍ iii) ی	i) ARABIC LETTER ALEF MAKSURA (YEH shaped letter with no dots at any position) ii) ARABIC LETTER YEH WITH TAIL (Pashto, Sindhi) iii) ARABIC LETTER FARSI YEH (Arabic, Persian, Urdu)

3. Conclusion

In Arabic script characters visually appear similar at different positions (initial, medial, final, and isolated). Presence of such characters causes confusions and they can be potential source of spoofing. Therefore confusing character’s shape problem has to be solved for secure implementation of Internationalized Domain Names (IDNs) and for Email Address Internationalization (EAI). In this document we identified confusing characters in Arabic script and we have divided these confusing characters into three categories. First category consist of those confusing characters that are at script level i.e. not distinct in any language. In second category those characters are identified that are confusing for at least one language but distinct in at least one language. Third category consists of those characters with similar shapes not distinct in any language.