Normalization Character List for Arabic Script Block

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

Revision History:

Date	Version	Changes	Changed by
04 Sep 2008	1.01	Created	Arfan Mansoor
26 May 2010	1.02	Reviewed	Rabia Sirhindi and Sarmad Hussain

Objective

In Arabic script block we have some characters that can be written in two forms: i) composed form and ii) decomposed form. In composed form characters occur as a single entity in Unicode block i.e. single codepoint is used to for that character e.g. آ (U+0622). In decomposed form the same character can be written by the combination two or more codepoints e.g. آ (U+0622) can be written by the combination of ا (U+0627) and ◌ٓ (U+0653). The formation of same character by more than one way can cause security risks because it is potential source of phishing problem in Internationalized Domain Names (IDNs).

The purpose of this document is to identify all such characters in Arabic script that can be formed in more than one way. Below is the list of these characters that exits in composed and decomposed forms in Arabic block. All such cases have to be normalized to make IDNs more secure.

Combining Mark	Composed Form	Decomposed Form	Unicode Normalized Form
◌ٓ U+0653	آ U+0622	ا ◌ٓ U+0627 U+0653	Defined
◌ٔ U+0654	أ U+0623	ا ◌ٔ◌ U+0627 U+0654	Defined
	ؤ U+0624	و ◌ٔ U+0648 U+0654	Defined
		ي ◌ٔ U+064A U+0654	Defined
		ى◌ٔ	Not Defined

Combining Mark	Composed Form	Decomposed Form	Unicode Normalized Form
	ئ U+0626	U+0649 U+0654	
		ی◌ٔ U+06CC U+0654	Not Defined
	ۀ U+06C0	ە◌ٔ U+06D5 U+0654	Defined
		ه◌ٔ U+0647 U+0654	Not Defined
	ۂ U+06C2	ہ◌ٔ U+06C1 U+0654	Defined
		ه◌ٔ U+0647 U+0654	Not Defined
	ۓ U+06D3	ے◌ٔ U+06D2 U+0654	Defined
	ځ U+0681	ح◌ٔ U+062D U+0654	Not Defined
	ݬ U+076C	ر◌ٔ U+0631 U+0654	Not Defined
◌ٕ U+0655	إ U+0625	ا◌ٕ U+0627 U+0655	Defined

Combining Mark	Composed Form	Decomposed Form	Unicode Normalized Form
◌ُ U+064F	ۇ U+06C7	و ◌ُ U+0648 U+064F	Not Defined
		و ◌ؙ U+0648 U+0619	Not Defined
◌ٰ U+0670	ۈ U+06C8	و ◌ٰ U+0648 U+0670	Not Defined
◌۬ U+06EC	ۏ U+06CF	و ◌۬ U+0648 U+06EC	Not Defined
	غ U+063A	ع ◌۬ U+0639 U+06EC	Not Defined
	ض U+0636	ص ◌۬ U+0635 U+06EC	Not Defined
	خ U+062E	ح ◌۬ U+062D U+06EC	Not Defined
	ڿ U+06BF	چ ◌۬ U+0686 U+06EC	Not Defined
	ذ U+0630	د ◌۬ U+062F U+06EC	Not Defined
	ز U+0632	ر ◌۬ U+0631 U+06EC	Not Defined
	ڶ U+06B6	ل ◌۬ U+0644 U+06EC	Not Defined
	ف	ٯ◌۬	Not Defined

Combining Mark	Composed Form	Decomposed Form	Unicode Normalized Form
	U+06A7	U+066F U+06EC	
	ف U+0641	ڡ ◌۬ U+06A1 U+06EC	Not Defined
	ن U+0646	ں ◌۬ U+06BA U+06EC	Not Defined
	ڬ U+06AC	ك ◌۬ U+0643 U+06EC	Not Defined
	ݢ U+0762	ک ◌۬ U+06A9 U+06EC	Not Defined
	ݥ U+0765	م ◌۬ U+0645 U+06EC	Not Defined
◌ؕ U+0615	ݲ U+0772	ح ◌ؕ U+062D U+0615	Not Defined
	ٹ U+0679	ٮ ◌ؕ U+066E U+0615	Not Defined
	ڑ U+0691	ر ◌ؕ U+0631 U+0615	Not Defined
	ڈ U+0688	د ◌ؕ U+062F U+0615	Not Defined
	ݱ U+0771	ڗ ◌ؕ U+0697 U+0615	Not Defined
	ڻ	ن ◌ؕ	Not Defined

Combining Mark	Composed Form	Decomposed Form	Unicode Normalized Form
	U+0768	U+0646 U+0615	
	ڋ U+068B	◌ؕ ڊ U+068A U+0615	Not Defined
	ڻ U+06BB	◌ؕ ں U+06BA U+0615	Not Defined
◌ٛ U+065B	ؽ U+063D	◌ٛ ی U+06CC U+065B	Not Defined
	ۉ U+06C9	◌ٛ و U+0648 U+065B	Not Defined
	ݾ U+077E	◌ٛ س U+0633 U+065B	Not Defined
	ۮ U+06EE	◌ٛ د U+062F U+065B	Not Defined
	ۯ U+06EF	◌ٛ ر U+0631 U+065B	Not Defined
	ۿ U+06FF	◌ٛ ھ U+06BE U+065B	Not Defined
		◌ٛ ه U+0647 U+065B	Not Defined
◌ۛ U+06DB	ؿ U+063F	◌ۛ ی U+06CC U+06DB	Not Defined
		◌ۛ ی	Not Defined

Combining Mark	Composed Form	Decomposed Form	Unicode Normalized Form
		U+0649 U+06DB	
	ش U+0634	س ◌ۛ U+0633 U+06DB	Not Defined
	ڜ U+069C	ڛ ◌ۛ U+069B U+06DB	Not Defined
	ث U+062B	ٮ◌ۛ U+066E U+06DB	Not Defined
	څ U+0685	ح ◌ۛ U+062D U+06DB	Not Defined
	ژ U+0698	ر ◌ۛ U+0631 U+06DB	Not Defined
	ڎ U+068E	د ◌ۛ U+062F U+06DB	Not Defined
	ڠ U+06A0	ع ◌ۛ U+0639 U+06DB	Not Defined
	ڤ U+06A4	ڡ ◌ۛ U+06A1 U+06DB	Not Defined
	ڨ U+06A8	ٯ◌ۛ U+066F U+06DB	Not Defined
	ڭ U+06AD	ك ◌ۛ U+0643 U+06DB	Not Defined
	ڴ U+06B4	گ ◌ۛ U+06AF U+06DB	Not Defined
	ڷ	ل ◌ۛ	Not Defined

Combining Mark	Composed Form	Decomposed Form	Unicode Normalized Form
	U+06B7	U+0644 U+06DB	
	ڽ U+06BD	ں ◌ۛ U+06BA U+06DB	Not Defined
	ݣ U+0763	ک ◌ۛ U+06A9 U+06DB	Not Defined
◌ٜ U+065C	ب U+0628	ٮ ◌ٜ U+066E U+065C	Not Defined
	ڊ U+068A	د ◌ٜ U+062F U+065C	Not Defined
	ڋ U+068B	ڈ ◌ٜ U+0688 U+065C	Not Defined
	ڔ U+0694	ر ◌ٜ U+0631 U+065C	Not Defined
	ڣ U+06A3	ف ◌ٜ U+0641 U+065C	Not Defined
	ڹ U+06B9	ن ◌ٜ U+0646 U+065C	Not Defined
	ۼ U+06FC	غ ◌ٜ U+063A U+065C	Not Defined
	ۻ U+06FB	ض ◌ٜ U+0636 U+065C	Not Defined
	ݑ U+0751	ث ◌ٜ U+062B U+065C	Not Defined

Combining Mark	Composed Form	Decomposed Form	Unicode Normalized Form
	ݦ U+0766	م ◌ٜ U+0645 U+065C	Not Defined
◌ٚ U+065A	ڵ U+06B5	ل ◌ٚ U+0644 U+065A	Not Defined
	ۆ U+06C6	و ◌ٚ U+0648 U+065A	Not Defined
	ێ U+06CE	ی ◌ٚ U+06CC U+065A	Not Defined
		ى ◌ٚ U+0649 U+065A	Not Defined
	ݖ U+0756	ٮ ◌ٚ U+066E U+065A	Not Defined
	ݩ U+0769	ن ◌ٚ U+0646 U+065A	Not Defined

In addition to these characters there are some characters that are formed by more than two characters combinations. These characters are listed in table given below.

ښ U+069A	س U+0633	◌ٜ U+065C	◌۬ U+06EC
ڣ U+06A3	ڡ U+06A1	◌ٜ U+065C	◌۬ U+06EC
ۺ U+06FA	س U+0633	◌ۛ U+06DB	◌ٜ U+065C
ۻ	ص	◌ٜ	◌۬

U+06FB	U+0635 U+065C U+06EC
ۼ U+06FC	ع ◌ٜ ◌۬ U+0639 U+065C U+06EC

Conclusion

Occurrence of characters in more than one form in Unicode Arabic script block can be a major security risk and these needs to be normalized to make IDNs more secure. In this document we have identified all those characters for Arabic script block that can occur in more than one form.