Subject: RE: Subj: security and collation in Arabic script -- hamza with aliph, tanween-al-fatah with aliph, and shaddah with vowel diacritic (was: tanween al fatah security/collation (follow up on earlier post))
From: CE Whitehead <cewcathar@hotmail.com>
Date: Mon, 26 Apr 2010 08:23:07 -0400
To: <rick@unicode.org>
CC: <ntounsi@gmail.com>, <ehsan@mozilla.com>

L2/10-127

Hi!
Attached are three test files (one showing the similarity between the kashmiri aliphs with hamzas and the 'traditional' -- hopefully these are compatibity characters at least; two showing typing order/display issues, one where the display of a consonant seat with a shaddah is confusable with a consonant display with a shaddah when the vowel diacritic is typed in the wrong order; and one where the logical typing order of the tanween-al-fatah varies -- as I think it should be allowed to do.
(thanks to Rick for letting me email this previously to the unicode email;
this gave me a chance to think all the issues through;
I should share this with the rest of the Arabic script experts on the unicode mail list for feedback --
but must first resubscribe to the mail list; in the meantime I cc'd two people from the bidi list; hope that is o.k.)

ISSUES

□$B!!□(B
I. IDN Question: Is a kashmiri aliph vowel seat with a hamza (U0672, U0673)
a compatibity equivalent of the more standard (I think) decomposable vowel seat with hamza? (see attached);
also, see: http://unicode.org/review/resolved-pri-100.html
U0673 was deprecated -- or will be;
I don't know how this affects IDN's; is this character still used in IDN's?
What about U0673? is that still in use?
Finally -- trivia I guess -- is the Saudi list of acceptable characters for idn's for Arabic the only list
(see http://www.iana.org/domains/idn-tables/tables/sa_ar_1.0.html)?
Is it used all over the Arabic world?
What list is used for India?
□$B!!□(B

II. Formatting/display issue
The disallowing of remapped compatibility aliph with tanween-al-fatah/fathatan sequence
(UFD3C; UFD3D) --
it seems that the vowel seats with hamza are allowed in idn's;
however I do not see other inflectional endings for Arabic in any of the legal IDN forms;
in the remapped compatibility forms (see http://unicode.org/reports/tr36/idn-chars.html) --
are all inflectional endings disallowed?
(what I can see with my browser is fuzzy anyway)
If so then UFD3C and UFD3D should not be allowed; otherwise I am unsure -- although these are confusable with
the remapped compatibility form for the Arabic aliph with the high hamza slightly to its right (in an rtl context; U0675).

□$B!!□(B
III. Typing Order and Internal network confusables (neither of these is an issue in idn security since these diacritics and also the decomposable form are disallowed in idn's)

-- tanween-al-fatah/fathatan diacritic typed logically before aliph and typed logically after aliph (at least in ie7, ie8; see attached)--

-- Consonant with shaddah versus consonant followed by vowel followed by shaddah (at least in ie7, ie8; see attached) --

□$B!!□(B
IV sorting/collation issues

-- tanween-al-fatah/fathatan diacritic typed logically before aliph and typed logically after aliph --
What is the best way to approach collating these two
(which I sort of feel should be compatibility equivalents -- I don't know what other Arabic speakers feelings are here
but I do think that the other logical typing order is possible)?
One option: when tanween-al-fatah occurs prior to an aliph within the same word insert a zero-spacing character
between the tanween-al-fatah and the aliph for collating purposes;
however -- is this done? is this sufficient?
(I wonder if these could be compatibility equivalents for the purposes of sorting
also for the purposes of internal network security?
Sorry to keep hammering at this topic.)

Best,
C. E. Whitehead
cewcathar@hotmail.com

> Date: Fri, 23 Apr 2010 15:46:58 -0700
> From: rick@unicode.org
> To: cewcathar@hotmail.com
> Subject: Re: Subj: tanween al fatah security/collation (follow up on earlier post)
>
> All right. You can e-mail me here... Feel free to send a document or
> whatever is needed, if more than plain text!
> Rick
>
>
> > [alas I only have one or two browsers so I may have to send a test
> file to you so I will use this email address too]).
> > I am sorry. I'll resubmit.
>

test_aliphwithhamzakashmirivsarabic.html

Aliph Vowel Seats with Hamzas

(a comparison of remapped compatibility forms, Kashmiri forms, and standard Arabic decomposable forms)
(NOTE: except for the vowel seats with hamzas all characters are atomic. The decomposable vowel seat with hamza always gives the proper display with the atomic characters. I have not checked the display in anything but ie8 at this point. It is possible that all of these are compatibility equivalents anyway and so then the issue is -- are these going to be normalized by all browsers?)

خطأ

(UFE84; Remapped compatibility aliph with hamza above, final form)

خطأ

(UFE83; Remapped compatibility aliph with hamza above, isolated form; the t.aa whether connecting or in isolation here looks about the same!)

خطأ

(U0672; Atomic aliph with hamza above, kashmiri -- note the display similarity to the other forms; in notepad as a text file however, this looks a bit like the remapped compatibility aliph with hamza, isolated form; however it should be noted that in Arabic it is not quite proper to use the isolated form in this context!)

خطأ

(U0623; The decomposable aliph with hamza above)

بالإجماع

(UFE84; Remapped compatibility aliph with hamza below, final form; this does not connect properly to the lam, which is a connecting letter and so the two should connect.)

بالإجماع

(UFE83; Remapped compatibity aliph with hamza below, isolated form; generally the isolated forms of the remapped compatibility characters do not cause confusion when logically following a connector; they simply do not connect properly so that the script does not look like Arabic.)

بالإجماع

(U0673; Atomic aliph with hamza below, Kashmiri; this should connect to the lam in this context, note the display similarity to the other forms; in notepad as a text file however, it does not connect like other connectors)

بالإجماع

(U0623; Decomposable aliph with hamza below)

إذا

(UFE83; Remapped compatibility aliph with hamza below, isolated form)

إذا

(U0673; Atomic aliph with hamza below, Kashmiri; note the display similarity to the other forms; it also looks like other forms in notepad as a text file)

إذا

(U0625; Decomposable aliph with hamza below)

أضاءت

(UFE83; Remapped compatibility aliph with hamza above, isolated form; the t.aa whether connecting or in isolation here looks about the same!)

أضاءت

(U0672; Atomic aliph with hamza above, kashmiri -- note the display similarity to the other forms; also in notepad as a text file, this looks quite a bit like the remapped compatibility aliph with hamza, isolated form and also like the decomposable form in this context)

أضاءت

(U0623; Decomposable aliph with hamza below)

test_tanweenalfatahandaliph.html

خبراً

Tanween-al-fatah (fathatan) -- U064B -- typed logically following the aliph -- U0627

خبرًا

Tanween-al-fatah (fathatan) -- U064b -- typed logically before the aliph -- U0627

test_vowelshaddahsequences.html

Arabic-consonant vowel shaddah (an illegal sequence in a normal Arabic word since the shaddah doubles the consonant; however canonicalization should put the shaddah first and the vowel diacritic second to my understanding, making this sequence fine). This sequence is rendered in IE7 and IE8 identically to Arabic-consonant shaddah with no vowel suggesting that canonicalization does not occur here prior to display (I understand both the shaddah and vowel diacritic to be combining marks and assume that the shaddah has a combining class closer to 0 than the vowel diacritic for proper display -- correct me if I am wrong); vowels in brackets [] are not actually encoded whereas vowels in brackets {} are encoded in the order shown -- which is the wrong order for ie7 and ie8:

f[a]-b[i]t-t[u]

فئتّ

confusable with
f[a]-b[i]t-{u}t

فئتَّ

but it should map to and display like:

فئتَّ

* * *

d[a]rr[a]s[a]

درّس

d[a]r{a}rs[a]

درَّس

but it should map to and display like:

درَّس

"Inadequate Rendering Support "An additional problem arises when a font and/or rendering engine has inadequate support for certain sequences of characters. These are characters that should be visually distinguishable, but don't appear that way. In example 8a, the a-umlaut is followed by another umlaut. The Unicode Standard guidelines indicate that the second umlaut should be 'stacked' above the first, producing a distinct visual difference. But as this example shows, common fonts will simply superimpose the second umlaut; and if the positioning is close enough, the user will not see a difference between 8a and 8b." There is no problem here in terms of the idn since the vowels are not available separately.

test_aliphwithhamzakashmirivsarabic.html	**Content-Type:** text/html

	Content-Encoding: base64

test_tanweenalfatahandaliph.html

test_tanweenalfatahandaliph.html	**Content-Type:** text/html **Content-Encoding:** base64

test_vowelshaddahsequences.html

test_vowelshaddahsequences.html	**Content-Type:** text/html **Content-Encoding:** base64