

Title: Application to include Arabic alphabet shapes to Arabic 0600 Unicode character set

Action: For consideration by UTC and ISO/IEC JTC1/SC2/WG2

Author: Mohammad Mohammad Khair

Date: 17-Dec-2018

Introduction:

Arabic language is spoken in all Arabic countries as well as taught in all Muslim countries around the world with a population of 1.9 Billion people. The Quran represents the religious holy text for this population and is written in Arabic language letters. Two of essential plain letter shapes which are variations of the Alf and Hamza shapes are missing from the Unicode Arabic 0600 table. Thus far, users have had to sequence multiple shapes of miscellaneous characters from the table Arabic 0600 to come up with these essential letter shapes as a composite, however, that has presented big issues for the accurate accounting of number of letters in the holy text. The proposed two new characters need to stand as their own due to each being an independent whole letter used in the Arabic language, especially in the Quran text. It is important they each get their own symbol in order for the Quran plain text character count to be equal to its symbol count that is making up the text. This is critically important as many new scientific and research publications are exploring structure and components of words, verses, and chapters in the Quran and their relationships, such as derived numeric statistics on the location and occurrence of letters and their frequency, which depend highly on the character count (i.e. symbol count) for each letter in each word, and their position within the text and verses. It is currently extremely difficult to conduct these studies due to the differences between the symbol count and the independent letter count, primarily because of the deficiency in the Unicode table from these two characters as independent letters, which we respectfully request the inclusion of these letter shapes or symbols in the Arabic 0600 table if possible, or its extension.

Detailed Description:

The Arabic language character set is in critical need to add the following two characters to the Arabic 0600 Unicode Character Set table for the following two symbols:

- 1) ل^ء
- 2) ء

These two symbols are very commonly used in the Islamic holy text of Quran throughout its verses. Each of these symbols represents a single letter in Arabic, i.e. counts as one letter, and should be represented in the Quran text as a single letter symbol, not made up of combined character sequence, which affects the letter counting of the holy text. Letter counting is essential to ensure the integrity of the holy Quran text, and therefore base letters can not be represented by a sequence of multiple Unicode symbols, as that presents a conflict when compared to the rest of the characters used in the Arabic language in the holy Quran text. Therefore there is an critical need to make these two symbols an indivisible part of the Arabic 0600 block so that they can be typed in as a single letter representation along with the rest of the characters of the Arabic language in the Quran text. Until today, users have had to type these two single letters symbols in Unicode by placing in sequence multiple symbols to make up the figure of these single letter representations such as using three symbols in sequence x0640 x0654 x0627 for the first letter and two symbols in sequence x0640 x0654 for the second letter, however, these are single letters and require their own single symbol in the Unicode 0600 Arabic block.

The field of numeric computational science of the Quran is an emerging new field of science that depends on accuracy of the letter count and representation in the digital text, and therefore having a complete character set in the Arabic 0600 table is essential for proper full representation of the Quran holy text in all its letter character shapes. This would to a great extent simplify the development of computations and statistics around the text due to its one to one mapping between its plain text (without superscripts or subscripts or signs) and the count of letter symbols used in that digital text. The current situation forcing a sequence of characters to represent a single letter leads to numerous letter counting issues when performing plain letter count of the holy text of the Quran, leading to inaccuracies in the letter count and other derived computations and statistics, due to the complexity of dealing with the sequence of Unicode symbols to represent a single letter. This is all due to the lack of representation of a single letter by a single symbol in the text using the Unicode table of characters. The addition of these two symbols will greatly facilitate the Quran numeric counting of letters and other key derived computations and statistics.

We propose using any currently available Unicode symbol values in the table Arabic 0600 such as 061D, and/or freeing up and using the Unicode values 0607 or 0606 for symbols representing shapes for mathematical third root and fourth root as they do not relate what so ever to Arabic language letters but rather are actually mathematical symbols and therefore do not belong to this table of Arabic 0600 which should be preserved for key Arabic language symbols of its letters and shapes variations.

The Quran text is the most published book in the world, therefore having sufficient symbols in Arabic Unicode table 0600 to adequately represent its essential base letters using one Unicode symbol per one letter symbol uniquely is of critical importance to enable the proper typing of its text representation, and its proper letter counting. We request the kind addition of the requested

two symbols described above and in the application submitted. Please forward any questions or inquiries to the following email:

mohammadkhair@gmail.com

USA +001-847-809-9090

Requests

The author requests the encoding 2 new alphabets used in Arabic language.

Character name and shape

Shape	Proposed Code Point	Name
ا	0606, or 061D	Hamza_Shahta_Alf
ء	0607	Hamza_Shahta

Example words with letter characters from the Quran text (www.quran.com) with chapter# : verse #

ا examples:	Quran Chapter: Verse Numbers	ء examples:	Quran Chapter: Verse Numbers
يَقَادُمُ	2:33	وَالصَّبِيِّينَ	2:62
بِأَيِّدِنَا	2:39	خُسَيْنَ	2:65
بِأَبْنِي	2:41	الَّذِينَ	2:71
بِأَيْتِ	2:61	خَطِيئَتُهُ	2:81
فَعَانَتْ	2:265	تَسْأَلُوا	2:108
بِإِخْذِيهِ	2:67	يَسْأَلُكَ	4:153
سَيِّئَاتِكُمْ	2:271		
الْمَعَابِ	3:14		

Notes on special behaviour cases of Letter ء

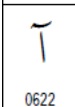
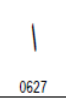
- The letter ء represents a single letter when present in the middle of the word, in other words, when it is followed by another letter.
- However, when ء occurs at the end of the word, meaning it is followed by a space

(hexadecimal value x0020 or x2000), then ء needs to actually be split to two letters symbols in sequence: ّ followed by ا, where ّ the is the new proposed letter symbol and ا is the x0627 Unicode letter Alif.

- In case the space at the end of the word is removed and replaced by another letter placed adjacent to ء then the two letters ّ followed by ا are rejoined to form single letter ء again.
- Example words where ء occurs at the end and needs to become two letters ّ followed by ا:



شيء, هنيا, مريا

- The letter can be combined with any superscript or subscript letters in the table Arabic 0600
- The character ء is considered to be a different style of writing for the first letter of the Arabic

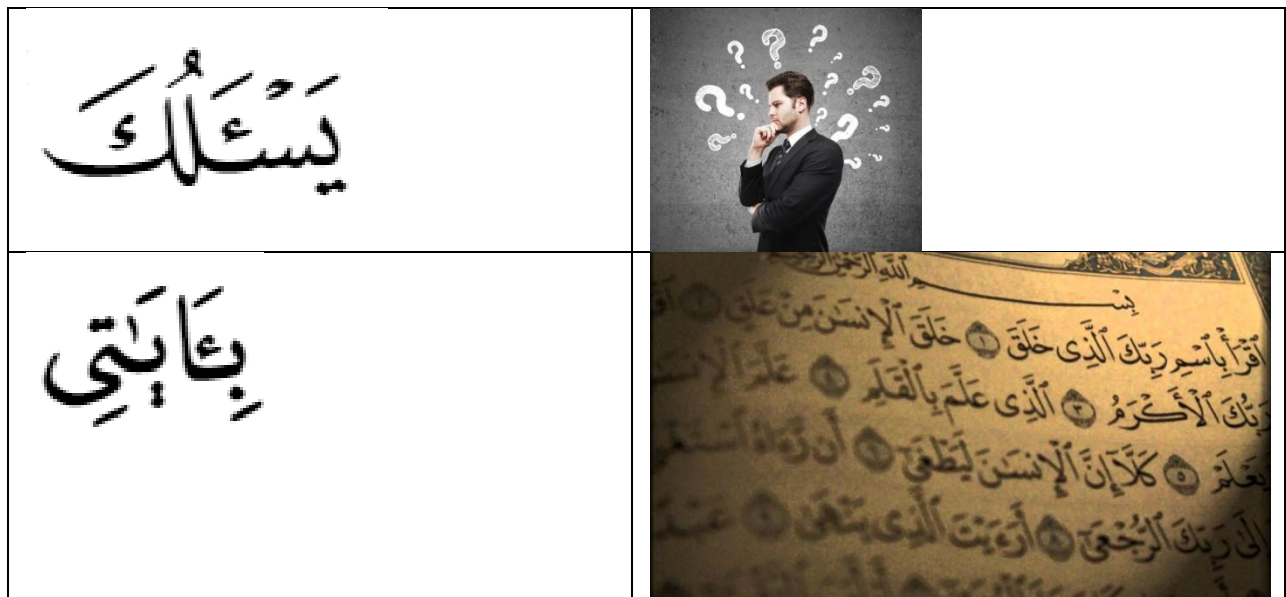
alphabet Alif Madda with unicode 0622, with corresponding shapes  0622 and it takes the same letter order precedence and value as Alif with unicode 0627  0627 the first character in the Arabic alphabet in during sorting and searching and indexing operations.

Notes on special behaviour cases of Letter ّ

- The letter ّ represents a single letter when typed, however, when it is followed by ا letter then the two are combined to form a single letter ء. The rules of behaviour for ء then.
- The letter can be combined with any superscript or subscript letters in the table Arabic 0600.

- The character ^ء is considered to be also a different style of writing the first letter of the Arabic alphabet Alif with unicode 0623, with corresponding shapes  and it takes the same letter order precedence and value as Alif with unicode 0627  the first character in the Arabic alphabet in during sorting and searching and indexing operations.

Figures:



References:

1. Quran www.quran.com
2. Quran text numeric letter and word count and derived computations
www.quranmetadata.com or <https://www.facebook.com/QuranMetaData>

ISO/IEC JTC 1/SC 2/WG 2

PROPOSAL SUMMARY FORM TO ACCOMPANY SUBMISSIONS

FOR ADDITIONS TO THE REPERTOIRE OF ISO/IEC 10646¹.

Please fill all the sections A, B and C below.

Please read Principles and Procedures Document (P & P) from <http://std.dkuug.dk/JTC1/SC2/WG2/docs/principles.html> for guidelines and details before filling this form.

Please ensure you are using the latest Form from <http://std.dkuug.dk/JTC1/SC2/WG2/docs/summaryform.html>.

See also <http://std.dkuug.dk/JTC1/SC2/WG2/docs/roadmaps.html> for latest *Roadmaps*.

A. Administrative

1. Title: **Addition of Three letters with Hamza Positions for Arabic Character Set**

2. Requester's name: **Mohammad Mohammad Khair**

3. Requester type (Member body/Liaison/Individual contribution): **Individual contribution**

4. Submission date: **November 12th, 2018**

5. Requester's reference (if applicable): **mohammadkhair@gmail.com**

6. Choose one of the following:

This is a complete proposal: **Yes**

(or) More information will be provided later:

B. Technical – General

1. Choose one of the following:

a. This proposal is for a new script (set of characters): **No**

Proposed name of script:

b. The proposal is for addition of character(s) to an existing block: **Yes**

Name of the existing block: **0600 Arabic**

2. Number of characters in proposal: **2**

3. Proposed category (select one from below - see section 2.2 of P&P document):

¹. Form number: N4502-F (Original 1994-10-14; Revised 1995-01, 1995-04, 1996-04, 1996-08, 1999-03, 2001-05, 2001-09, 2003-11, 2005-01, 2005-09, 2005-10, 2007-03, 2008-05, 2009-11, 2011-03, 2012-01)

Application to include Arabic alphabet shapes to Arabic 0600 Unicode character set

A-Contemporary

Y

B.1-Specialized (small collection)

B.2-Specialized (large collection)

C-Major extinct

D-Attested extinct

E-Minor extinct

F-Archaic Hieroglyphic or Ideographic

G-Obscure or questionable usage symbols

4. Is a repertoire including character names provided?

Yes

a. If YES, are the names in accordance with the "character naming guidelines"

in Annex L of P&P document?

b. Are the character shapes attached in a legible form suitable for review?

Yes

5. Fonts related:

a. Who will provide the appropriate computerized font to the Project Editor of 10646 for publishing the standard?

Mohammad Mohammad Khair mohammadkhair@gmail.com

b. Identify the party granting a license for use of the font by the editors (include address, e-mail, ftp-site, etc.):

Mohammad Mohammad Khair mohammadkhair@gmail.com

6. References:

a. Are references (to other character sets, dictionaries, descriptive texts etc.) provided?

Yes

b. Are published examples of use (such as samples from newspapers, magazines, or other sources)

of proposed characters attached?

Yes, also reference **Quran.com** for example Quran verses include (chapter# : verse#) listed below:

أ

examples:

يَقَادُمُ

2:33

بِأَيَّتِنَا

2:39

بِأَيَّتِي

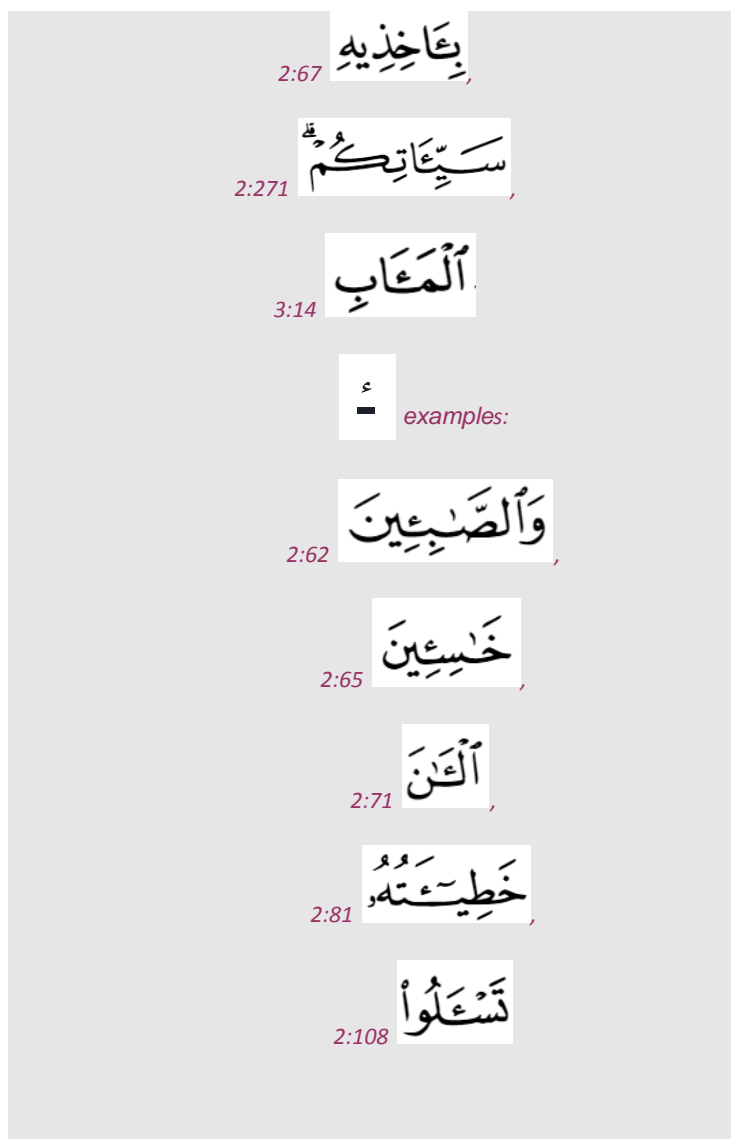
2:41

بِأَيَّتِ

2:61

فَعَانَتْ


2:265




7. Special encoding issues:






Does the proposal address other aspects of character data processing (if applicable) such as input, presentation, sorting, searching, indexing, transliteration etc. (if yes please enclose information)?

Yes

The character  is considered to be a different style of writing for the first letter of the Arabic alphabet Alif Madda







with unicode 0622, with corresponding shapes  0622 and it takes the same letter order precedence and value as Alif





 0627 the first character in the Arabic alphabet in during sorting and searching and indexing operations.






<p>The character  is considered to be also a different style of writing the first letter of the Arabic alphabet Alif with</p> <p>unicode 0622, with corresponding shapes  and it takes the same letter order precedence and value as Alif</p> <p> the first character in the Arabic alphabet in during sorting and searching and indexing operations.</p> <p>See the sections titled Character Name and Shape as well as Notes for behaviour for  and .</p>
<p>8. Additional Information:</p> <p>Submitters are invited to provide any additional information about Properties of the proposed Character(s) or Script that will assist in correct understanding of and correct linguistic processing of the proposed character(s) or script. Examples of such properties are: Casing information, Numeric information, Currency information, Display behaviour information such as line breaks, widths etc., Combining behaviour, Spacing behaviour, Directional behaviour, Default Collation behaviour, relevance in Mark Up contexts, Compatibility equivalence and other Unicode normalization related information. See the Unicode standard at http://www.unicode.org for such information on other scripts. Also see Unicode Character Database (http://www.unicode.org/reports/tr44/) and associated Unicode Technical Reports for information needed for consideration by the Unicode Technical Committee for inclusion in the Unicode Standard.</p>

C. Technical - Justification

1. Has this proposal for addition of character(s) been submitted before?	No
If YES explain	
2. Has contact been made to members of the user community (for example: National Body, user groups of the script or characters, other experts, etc.)?	Yes
If YES, with whom?	Dr. Khaled Bakro, International University Of Renewal, dr.khaled.bakro@gmail.com http://www.tajdeeduniversity.com/
If YES, available relevant documents:	Letter Of Support for Character Addition to Unicode Arabic 0600 set, Quran text containing requested characters to be added Quran.com
3. Information on the user community for the proposed characters (for example: size, demographics, information technology use, or publishing use) is included?	Yes
Reference:	Quran text is most published text in the world, used by Muslim population of 1.9 Billion

4. The context of use for the proposed characters (type of use; common or rare)		Common
Reference:	<i>To be used for accurate representation of Quran characters as unique character symbols for accurate letter counting and other numeric computations and statistics per letter.</i>	
5. Are the proposed characters in current use by the user community?		Yes
If YES, where? Reference:	 <p><i>Proposed characters  and  are in wide use, however they have to be created by sequencing multiple (two or three) parts or unicode symbols instead of a single symbol representing a single letter which is critical for letter counting in the Quran holy text and to avoid addition of letter symbols not in equivalence to its letter count.</i></p>	
6. After giving due considerations to the principles in the P&P document must the proposed characters be entirely in the BMP?		Yes
If YES, is a rationale provided?		Yes
If YES, reference:	 <p><i>Character symbols requested  and  are essential representation of base letter shapes used in the Quran text, used very commonly within the holy text, and each counts as a single letter. Each unique letter symbol therefore requires a single unique Unicode representation. Therefore it is critical that the two characters be co-located with the other characters set in the 0600 Arabic block of character symbols. Preferably together for ease of recognition by users if possible. Each of these two symbols counts as a single letter in Arabic and should be represented in the Quran text as a single letter. Until today, users have had to type these two single letters symbols in Unicode by placing in sequence multiple symbols to make up the figure of these single letter representations such as using three symbols x0640 x0654 x0627 for the first letter and x0640 x0654 for the second letter, however, these are single letters and require their own single symbol in the Unicode 0600 Arabic block. We propose using any currently available Unicode symbol values in the table Arabic 0600 such as 061D, and/or freeing up and using the Unicode values 0607 or 0606 for symbols representing square root shapes as they do not represent any Arabic letters but are actually mathematical symbols and therefore do not belong to this table of Arabic 0600.</i></p>	
7. Should the proposed characters be kept together in a contiguous range (rather than being scattered)?		Prefer
8. Can any of the proposed characters be considered a presentation form of an existing character or character sequence?		No
If YES, is a rationale for its inclusion provided?		<i>Each symbol is its own unique letter</i>

		presentation
If YES, reference:		
9. Can any of the proposed characters be encoded using a composed character sequence of either		
existing characters or other proposed characters?		Yes
If YES, is a rationale for its inclusion provided?		
If YES, reference:		
<p>Character symbols requested  and  are essential representation of base letter shapes used in the Quran text, used very commonly within the holy text, and each counts as a single letter. Each unique letter symbol therefore requires a single unique Unicode representation. Therefore it is critical that the two characters be co-located with the other characters set in the 0600 Arabic block of character symbols. Preferably together for ease of recognition by users if possible. Each of these two symbols counts as a single letter in Arabic and should be represented in the Quran text as a single letter. Until today, users have had to type these two single letters symbols in Unicode by placing in sequence multiple symbols to make up the figure of these single letter representations such as using three symbols x0640 x0654 x0627 for the first letter and x0640 x0654 for the second letter, however, these are single letters and require their own single symbol in the Unicode 0600 Arabic block. We propose using any currently available Unicode symbol values in the table Arabic 0600 such as 061D, and/or freeing up and using the Unicode values 0607 or 0606 for symbols representing square root shapes as they do not represent any Arabic letters but are actually mathematical symbols and therefore do not belong to this table of Arabic 0600.</p>		
10. Can any of the proposed character(s) be considered to be similar (in appearance or function)		
to, or could be confused with, an existing character?		No
If YES, is a rationale for its inclusion provided?		<p> and </p> <p>present unique shapes for letters used in the Quran. No confusion</p>
If YES, reference:		
11. Does the proposal include use of combining characters and/or use of composite sequences?		Yes
If YES, is a rationale for such use provided?		

<p>If YES, reference:</p>	<p>Characters proposed  and  can be combined in sequence with Any of the superscript or subscript symbols in the table 0600 Arabic. The combined sequential superscript or subscript should appear on top of the part  shape for both letters</p>
<p>Is a list of composite sequences and their corresponding glyph images (graphic symbols) provided?</p>	
<p>If YES, reference:</p>	<p>Yes for both  and </p>
<p>12. Does the proposal contain characters with any special properties such as</p>	
<p>control function or similar semantics?</p>	<p>No</p>
<p>If YES, describe in detail (include attachment if necessary)</p>	
<p></p>	
<p>13. Does the proposal contain any Ideographic compatibility characters?</p>	
<p>If YES, are the equivalent corresponding unified ideographic characters identified?</p>	<p>No</p>
<p>If YES, reference:</p>	<p></p>