

Proposal to encode productive Arabic-script modifier marks

Date: May 16, 2003
 Author: Jonathan Kew, SIL International Kamal Mansour, Agfa Monotype Mark Davis, IBM
 Address: Horsleys Green Agfa Monotype
 High Wycombe 425 Sherman Ave
 Bucks HP14 3XL Palo Alto, CA 94306
 England USA
 Tel: +44 (1494) 682306 +1 (650) 321-4102
 Email: jonathan_kew@sil.org kamal.mansour@agfamonotype.com mark.davis@jtcsv.com

A. Administrative

1. Title	Proposal to encode productive Arabic-script modifier marks
2. Requester's name	SIL International (contacts: Jonathan Kew, Peter Constable); Agfa Monotype (contact: Kamal Mansour)
3. Requester type	Expert contribution
4. Submission date	May 16, 2003
5. Requester's reference	
6a. Completion	This is a complete proposal.
6b. More information to be provided?	Only as required for clarification.

B. Technical — General

1a. New script? Name?	No
1b. Addition of characters to existing block? Name?	Yes — Arabic
2. Number of characters in proposal	23
3. Proposed category	A
4. Proposed level of implementation and rationale	2 (includes combining marks)
5a. Character names included in proposal?	Yes
5b. Character names in accordance with guidelines?	Yes
5c. Character shapes reviewable?	Yes
6a. Who will provide computerized font?	Jonathan Kew, SIL International
6b. Font currently available?	Yes
6c. Font format?	TrueType
7a. Are references (to other character sets, dictionaries, descriptive texts, etc.) provided?	See the documents cited under References.
7b. Are published examples (such as samples from newspapers, magazines, or other sources) of use of proposed characters attached?	See the documents cited under References.
8. Does the proposal address other aspects of character data processing?	Yes, suggested character properties and collation weights are included.

C. Technical – Justification

1.	Has this proposal for addition of character(s) been submitted before?	No (but see L2/02-021 for a related proposal).
2a.	Has contact been made to members of the user community?	Yes
2b.	With whom?	Academics working with Arabic script; language communities in South Asia and North Africa.
3.	Information on the user community for the proposed characters is included?	Yes
4.	The context of use for the proposed characters	Current use in publications in languages of North Africa, South Asia, and Near & Middle East; experimental orthographies for previously unwritten minority languages; scholarly and pedagogical use.
5.	Are the proposed characters in current use by the user community?	Yes
6a.	Must the proposed characters be entirely in the BMP?	Yes
6b.	Rationale?	Contemporary characters in current use
7.	Should the proposed characters be kept together in a contiguous range?	Preferably (for convenience of users and implementers), but not essential.
8a.	Can any of the proposed characters be considered a presentation form of an existing character or character sequence?	No
8b.	Rationale for inclusion?	N/A
9a.	Can any of the proposed characters be considered to be similar (in appearance or function) to an existing character?	Several proposed characters (#1, #2, #8-#10, #19-#22) appear similar to combining marks in the U+03xx block. Character #7 looks similar to U+0615. Character #10 looks similar to U+0652. Character #23 looks similar to U+06BA.
9b.	Rationale for inclusion?	Characteristic appearance of dots used in Arabic script differs from generic combining marks; different combining classes needed from U+03xx marks. U+0615 is a higher-level (phrasal) mark, and U+0652 is a diacritic that operates at the level of the vowel marks; neither is a component used in creating new letters. #23 differs from U+06BA in joining behavior.
10.	Does the proposal include the use of combining characters and/or use of composite sequences?	Yes, all but one of the proposed characters are combining
11.	Does the proposal contain characters with any special properties?	Combining marks

D. SC2/WG2 Administrative

To be completed by SC2/WG2

1. Relevant SC2/WG2 document numbers
2. Status (list of meeting number and corresponding action or disposition)
3. Additional contact to user communities, liaison organizations, etc.
4. Assigned category and assigned priority/time frame

Other comments

I. Background

The Universal Character Set currently encodes Arabic letters as indivisible units. However, the structure of the script is better understood as a small set of “skeletal” letterforms, to which modifier marks (primarily patterns of dots, but other marks are also used) are added to differentiate additional sounds (or letters) as needed to write a particular language. The creation of specific *base letter + modifier* combinations is a living, productive process, still in use today as Arabic script is adopted for writing additional languages.

To more adequately model and support the productive nature of modifier mark usage in Arabic, it is proposed that a set of combining modifier characters be added to the UCS Arabic repertoire.

In addition to allowing users to encode any combination of skeletal Arabic letterform plus modifiers, supporting the real-world evolutionary use of the script in transcribing new languages, this model also enables scholars to encode historical documents where modifiers may have been added over time, and allows the modifiers to be encoded individually where required (e.g., in pedagogical materials). In these, it may be necessary to apply markup separately to base characters and modifiers; this cannot generally be done using only precomposed letters.

While supporting a set of combining Arabic modifiers does require some effort on the part of implementers, especially of rendering systems and fonts, it is not substantially different in nature or complexity from that already required to properly support the Arabic vowel diacritics, Koranic marks, etc. Moreover, both standardization bodies and implementers will benefit from the fact that the Arabic block can be stabilized more quickly in this way than by encoding each new combination of *base letter + modifiers* in its own right, as and when it is proposed and documented as an individual new character.












II. Proposal













1. Repertoire

The following list of characters are proposed as additions to the UCS; all are combining marks, with the exception of #23.

For each modifier mark, one or more examples are given of letters where the mark has been used as a modifier on an undotted skeletal Arabic letterform. Where possible, examples of such “precomposed” letters from the current UCS repertoire are included. In addition, for most of the marks there are examples showing how they have been used to create new letters for minority languages whose orthographies are not yet well standardized or documented.

The letters and languages mentioned here are merely representative of the widespread practice of “extending” the Arabic script in this manner; additional examples can be found in the references listed at the end of this document.

- | | | | |
|-----|---|---|---|
| 1. |  | ARABIC MODIFIER SINGLE DOT ABOVE | خ U+062E, ض U+0636, م (older Songhoy, Fulfulde) |
| 2. |  | ARABIC MODIFIER TWO DOTS HORIZONTALLY ABOVE | ة U+0629, ت U+062A, ح (Songhoy, Fulfulde) |
| 3. |  | ARABIC MODIFIER TWO DOTS VERTICALLY ABOVE | ث U+067A, خ U+0682, ش (Gawri) |
| 4. |  | ARABIC MODIFIER THREE DOTS UPWARDS ABOVE | ث U+062B, ش U+0634, خ (Wolof, Fulfulde) |
| 5. |  | ARABIC MODIFIER THREE DOTS DOWNWARDS ABOVE | ت U+067D, ذ U+068F, غ (Hausa) |
| 6. |  | ARABIC MODIFIER FOUR DOTS ABOVE | ث U+067F, ذ U+0690, ش (Shina) |
| 7. |  | ARABIC MODIFIER TAH ABOVE | ث U+0679, ذ U+0688, ن (Saraiki, Shina) |
| 8. |  | ARABIC MODIFIER SMALL V ABOVE | ز U+0692, ل U+06B5, ن (Gojri) |
| 9. |  | ARABIC MODIFIER INVERTED V ABOVE | ؤ U+06C9, ذ U+06EE |
| 10. |  | ARABIC MODIFIER RING ABOVE | و U+06C4, ن (Punjabi) |
| 11. |  | ARABIC MODIFIER SINGLE DOT BELOW | ب U+0628, ج U+062C, م (Maba) |

12.		ARABIC MODIFIER TWO DOTS HORIZONTALLY BELOW	ي U+064A, ج U+0683, ن (Songhoy, Fulfulde)
13.		ARABIC MODIFIER TWO DOTS VERTICALLY BELOW	ب U+067B, ج U+0684
14.		ARABIC MODIFIER THREE DOTS UPWARDS BELOW	پ (Fulfulde)
15.		ARABIC MODIFIER THREE DOTS DOWNWARDS BELOW	پ U+067E, ج U+0686, ك (Kalasha)
16.		ARABIC MODIFIER THREE DOTS HORIZONTALLY BELOW	پ (Fulfulde)
17.		ARABIC MODIFIER FOUR DOTS BELOW	پ U+0680, ج U+0687
18.		ARABIC MODIFIER TAH BELOW	ج (Khowar)
19.		ARABIC MODIFIER SMALL V BELOW	ر U+0695
20.		ARABIC MODIFIER INVERTED V BELOW	پ (Songhoy, Fulfulde)
21.		ARABIC MODIFIER RING BELOW	ت U+067C, ظ U+0689
22.		ARABIC MODIFIER BAR THROUGH LETTER	و U+06C5, ل (Marwari)
23.		ARABIC LETTER DOTLESS NOON	ن (Punjabi)

At the time of writing, there are fewer than 30 unallocated positions in the Arabic block at U+06xx; with 23 new characters in the current proposal, and keeping in mind that there are known to be a number of additional vowels and other diacritics still to be encoded, it seems clear that the 06xx block is going to overflow. It is suggested, therefore, that ARABIC LETTER DOTLESS NOON be encoded in the existing Arabic block (perhaps at 065F), but that a separate extension block be allocated for the modifier marks. For the convenience of implementers, it would be helpful for this to be close to the 06xx block. One possibility would be to move one of the tentatively-allocated scripts in the Roadmap to space higher in the BMP, to allow creation of an Arabic Modifiers block.

2. Properties

All the proposed characters except for #23 ARABIC LETTER DOTLESS NOON are non-spacing combining marks, with General Category Mn and bidirectional type NSM. They have no numeric values, are not mirrored, and have no decompositions.

It has been suggested that they should all have combining class 0, and therefore would not participate in canonical reordering of combining characters. This is attractive in its simplicity, and allows users to encode any sequence knowing that no process will modify the ordering of the modifier marks. However, it has the disadvantage that where modifiers are used both above and below a single base letter, there would be multiple visually indistinguishable sequences and canonical ordering would not address this situation.

It is therefore proposed that these new Arabic combining marks be given new, low-value combining classes. (They should not use the existing classes in the 200s, as this would put them further from the base letter than vowel points during canonical ordering, greatly complicating rendering and other processes.) The assignments proposed are:

- Class 1 (already used for combining overlays): #22 ARABIC MODIFIER BAR THROUGH LETTER.
- Class 2: Arabic modifier marks below the base character (#11-#21).
- Class 3: Arabic modifier marks above the base character (#1-#10).

The one base character proposed here, ARABIC LETTER DOTLESS NOON, has General Category Lo, combining class 0, bidirectional type AL, and no other special properties. Its Arabic joining class is NOON. It differs from ن U+06BA ARABIC LETTER NOON GHUNNA in that U+06BA in initial and medial forms acquires a single dot above; the proposed ARABIC LETTER DOTLESS NOON, being intended as a base for the Arabic modifier marks, should not exhibit this behavior.

3. Usage notes and normalization issues

As a general rule, a skeletal (“dotless”) Arabic letter may be modified by the addition of one or more of these modifier marks, to form a new extended Arabic-script consonant (or occasionally a vowel, in some

languages). Marks may be used both above and below the same letter (as illustrated by some of the precomposed letters already in Unicode); it is also permissible to stack multiple marks either above or below, as shown in a few of the examples above.

Implementers (in particular, font developers) should note that certain combinations of *base letter + modifier* may require special positioning. This can be seen in a number of the extended Arabic characters already encoded, such as ٶ U+0696, where the position of the ARABIC MODIFIER SINGLE DOT ABOVE is different from that seen on ٷ U+0632. Another example is the use of the ARABIC MODIFIER RING BELOW, typically written attached to the base letter as in Pashto ټ U+067C. The ARABIC MODIFIER RING ABOVE is seen attached to the base letter in ٺ U+06C4, but is written separately in Punjabi ٺ (in one of several competing orthographies).

With the encoding of a set of Arabic modifier marks, it will be possible to represent many Arabic-script letters in two ways: either using the existing precomposed letters available in the U+06xx block, or using dotless letters plus modifier marks. As the existing precomposed letters have no decompositions (and stability policies prohibit adding them), the two forms will not be canonically equivalent. This means that there will be multiple “spellings” possible for Arabic script text, and Unicode normalization processes will not treat them as equivalent. It is strongly recommended that the existing precomposed letters should be used wherever possible, with sequences involving the new modifier marks being used *only* to encode letters that are not otherwise available.

An additional twist to the question of precomposed versus decomposed representations is that it would be possible to add modifier marks to precomposed letters that already incorporate modifiers of their own, such as adding the mark #4 ٲ ARABIC MODIFIER THREE DOTS UPWARDS ABOVE to ٺ U+062C ARABIC LETTER JEEM to form the Wolof letter ٲ. (Alternatively, this same letter could be created with mark #11 ٲ ARABIC MODIFIER SINGLE DOT BELOW applied to ٲ U+0685.) In the absence of canonical equivalence between the precomposed and decomposed forms, these would represent different “spellings” of the same extended Arabic letter. However, it is strongly recommended that users avoid such mixtures of precomposed and decomposed representation, and always use a fully-decomposed sequence for any letter that is not directly available as a precomposed form.

A data file, *DiscArabCombSeq.txt*, is provided (see Appendix A) that lists all the skeletal Arabic letters to which the modifier marks may be added. Use of these marks with other base characters, although not illegal, is considered non-standard practice and is unlikely to be well supported by fonts and rendering systems. The file also lists the sequences of skeletal letters and modifier marks that are strongly discouraged because they should be visually indistinguishable from existing Arabic letters. Arabic text processing systems may wish to offer users the option to convert between these sequences and the corresponding precomposed letters, or otherwise treat them as equivalent for certain purposes.

Implementers can encourage usage in accordance with these recommendations through the keyboard layouts or other input methods that are provided for languages with letters that must be composed using sequences. Implementations could even provide options to detect and warn users if these marks are applied to base characters other than the expected skeletal Arabic letters, and if sequences listed as “discouraged” in *DiscArabCombSeq.txt* are found. (For example, rendering systems could have the ability to display such combinations in visibly distinct ways, such as with marks serialized instead of stacked.)

4. Collation

For the purposes of the Unicode Collation Algorithm, it is proposed that the modifier marks be treated as Level 1 ignorables in the Default Unicode Collation Element Table, and given Level 2 weights. This means that all letters based on the same skeletal form will sort together in the default ordering at the primary level.

However, it is assumed that for any language where a new extended Arabic letter, encoded using a sequence of *base + modifier marks*, is used as part of the alphabet, the collation sequence would be tailored to sort the specific sequences used into their proper alphabetical positions.

A data file, *ArabicModifierKeys.txt*, is provided (see Appendix B) that lists proposed weights for the modifier marks, designed for use as an extension to version 3.1.1 of the standard *allkeys.txt* file. Note that ARABIC LETTER DOTLESS NOON should be inserted with a Level 1 weight immediately following U+0646 ARABIC LETTER NOON, affecting all following Level 1 weights. This will be only one of a number of changes with global effect required for a complete UCA update; the data given here illustrates how the new modifier marks would be integrated but does not represent a complete update to the Default Unicode Collation Element Table.

IV. Names list

Addition to Arabic block

065F ﺝ ARABIC LETTER DOTLESS NOON

Arabic modifier marks above

xx00 ◌̇ ARABIC MODIFIER SINGLE DOT ABOVE
xx01 ◌̈ ARABIC MODIFIER TWO DOTS HORIZONTALLY ABOVE
xx02 ◌̈́ ARABIC MODIFIER TWO DOTS VERTICALLY ABOVE
xx03 ◌̈́ ARABIC MODIFIER THREE DOTS UPWARDS ABOVE
xx04 ◌̈́ ARABIC MODIFIER THREE DOTS DOWNWARDS ABOVE
xx05 <reserved>
xx06 ◌̈́ ARABIC MODIFIER FOUR DOTS ABOVE
xx07 ◌̈́ ARABIC MODIFIER TAH ABOVE
xx08 ◌̈́ ARABIC MODIFIER SMALL V ABOVE
xx09 ◌̈́ ARABIC MODIFIER INVERTED V ABOVE
xx0A ◌̈́ ARABIC MODIFIER RING ABOVE
xx0B <reserved>
xx0C <reserved>
xx0D <reserved>
xx0E <reserved>
xx0F <reserved>

Arabic modifier marks below

xx10 ◌̇ ARABIC MODIFIER SINGLE DOT BELOW
xx11 ◌̈ ARABIC MODIFIER TWO DOTS HORIZONTALLY BELOW
xx12 ◌̈́ ARABIC MODIFIER TWO DOTS VERTICALLY BELOW
xx13 ◌̈́ ARABIC MODIFIER THREE DOTS UPWARDS BELOW
xx14 ◌̈́ ARABIC MODIFIER THREE DOTS DOWNWARDS BELOW
xx15 ◌̈́ ARABIC MODIFIER THREE DOTS HORIZONTALLY BELOW
xx16 ◌̈́ ARABIC MODIFIER FOUR DOTS BELOW
xx17 ◌̈́ ARABIC MODIFIER TAH BELOW
xx18 ◌̈́ ARABIC MODIFIER SMALL V BELOW
xx19 ◌̈́ ARABIC MODIFIER INVERTED V BELOW
xx1A ◌̈́ ARABIC MODIFIER RING BELOW
xx1B <reserved>
xx1C <reserved>
xx1D <reserved>
xx1E <reserved>
xx1F <reserved>

Arabic modifier marks through

xx20 ◌̈́ ARABIC MODIFIER BAR THROUGH LETTER

V. References

- Baart, Joan L. G. and Muhammad Zaman Sagar. 2002. *The Gawri language of Kalam and Dir Kohistan*.
(http://www.geocities.com/kcs_kalam/gawri.pdf)
- Chtatou, Mohamed. 1992. *Using Arabic script in writing the languages of the peoples of Muslim Africa*. Rabat:
Institute of African Studies.
- Davis, Mark and Kamal Mansour. 2002. *Proposal to amend Arabic repertoire*. L2/02-021.
- Kew, Jonathan. 2002. *Proposal for extensions to the Arabic block*. L2/02-274.
- . 2002. *Encoding generative Arabic nuktas: normalization concerns*.
(<http://www.jfkew.plus.com/unicode/Generative-Arabic.pdf>)
- . 2003. *Encoding Arabic extensions: options for the future of Unicode*. L2/03-044.
- . 2003. *Images of potential extended Arabic characters*. L2/03-051.
- Mansour, Kamal. 2002. *Progress report on L2/02-21 (Proposal to amend Arabic repertoire)*. L2/02-161.
- . 2003. http://www.bisharat.net/A12N/Afro-Arabic_Symbols.pdf.

Appendix A: Expected base characters & discouraged sequences

The machine-readable data file *DiscArabCombSeq.txt* provides a list of all the skeletal Arabic letters that are normally expected to be used as base characters for the modifier marks. Then it also lists the specific sequences beginning with these base characters that are *discouraged* in normal use, as they represent letters that are already encoded as individual Unicode characters.

The draft data file listed here uses abbreviated names to refer to proposed characters for which Unicode codepoints are not yet available.

Listing of data file *DiscArabCombSeq.txt*

```
# This file documents "expected" and "discouraged" usage for the Arabic-script modifier marks.
#
# There are three sections to the file.
#
# First, there is a list of the characters classified as "Arabic-script modifier marks".
#
# Second, a list of all the characters that are considered normal bases for the modifier marks;
# using any of the modifier marks on base characters not in this list is considered non-standard
# and is strongly discouraged unless there is a clear need and no alternative code sequence.
#
# Third, there is a list of sequences involving an "expected" base plus one or more
# modifier marks that are strongly discouraged, because they would be visually identical to
# pre-existing Arabic script characters. Such sequences should never be interchanged.
#
# On each line, there are two data fields separated by semicolon. Field 1 gives one or more
# Unicode scalar values; field 2 contains a code indicating which of the three types of code
# or sequence is represented by field 1.
#
# Field 2 format:
# M: arabic modifier marks
# B: normal base character for application of arabic modifier marks
# X: sequences that are strongly discouraged
#
# Characters classified as Arabic-script modifier marks:
#
<one dot above> ; M # ARABIC MODIFIER SINGLE DOT ABOVE
<two dots horiz above> ; M # ARABIC MODIFIER TWO DOTS HORIZONTALLY ABOVE
<two dots vert above> ; M # ARABIC MODIFIER TWO DOTS VERTICALLY ABOVE
<three dots up above> ; M # ARABIC MODIFIER THREE DOTS UPWARDS ABOVE
<three dots down above> ; M # ARABIC MODIFIER THREE DOTS DOWNWARDS ABOVE
<four dots above> ; M # ARABIC MODIFIER FOUR DOTS ABOVE
<tah above> ; M # ARABIC MODIFIER TAH ABOVE
<small v above> ; M # ARABIC MODIFIER SMALL V ABOVE
<inverted v above> ; M # ARABIC MODIFIER INVERTED V ABOVE
<ring above> ; M # ARABIC MODIFIER RING ABOVE
<one dot below> ; M # ARABIC MODIFIER SINGLE DOT BELOW
<two dots horiz below> ; M # ARABIC MODIFIER TWO DOTS HORIZONTALLY BELOW
<two dots vert below> ; M # ARABIC MODIFIER TWO DOTS VERTICALLY BELOW
<three dots up below> ; M # ARABIC MODIFIER THREE DOTS UPWARDS BELOW
<three dots down below> ; M # ARABIC MODIFIER THREE DOTS DOWNWARDS BELOW
<three dots horiz below> ; M # ARABIC MODIFIER THREE DOTS HORIZONTALLY BELOW
<four dots below> ; M # ARABIC MODIFIER FOUR DOTS BELOW
<tah below> ; M # ARABIC MODIFIER TAH BELOW
<small v below> ; M # ARABIC MODIFIER SMALL V BELOW
<inverted v below> ; M # ARABIC MODIFIER INVERTED V BELOW
<ring below> ; M # ARABIC MODIFIER RING BELOW
<bar through> ; M # ARABIC MODIFIER BAR THROUGH LETTER
#
# Characters considered "expected" base characters for the Arabic-script modifier marks:
#
0621 ; B # ARABIC LETTER HAMZA
0627 ; B # ARABIC LETTER ALEF
062D ; B # ARABIC LETTER HAH
062F ; B # ARABIC LETTER DAL
0631 ; B # ARABIC LETTER REH
0633 ; B # ARABIC LETTER SEEN
0635 ; B # ARABIC LETTER SAD
0637 ; B # ARABIC LETTER TAH
0639 ; B # ARABIC LETTER AIN
0643 ; B # ARABIC LETTER KAF
0644 ; B # ARABIC LETTER LAM
0645 ; B # ARABIC LETTER MEEM
```

```

0647 ; B # ARABIC LETTER HEH
0648 ; B # ARABIC LETTER WAW
0649 ; B # ARABIC LETTER ALEF MAKSURA
066E ; B # ARABIC LETTER DOTLESS BEH
066F ; B # ARABIC LETTER DOTLESS QAF
06A1 ; B # ARABIC LETTER DOTLESS FEH
06A9 ; B # ARABIC LETTER KEHEH
06AA ; B # ARABIC LETTER SWASH KAF
06AF ; B # ARABIC LETTER GAF
06BE ; B # ARABIC LETTER HEH DOACHASHMEE
06C1 ; B # ARABIC LETTER HEH GOAL
06CC ; B # ARABIC LETTER FARSI YEH
06CD ; B # ARABIC LETTER YEH WITH TAIL
06D2 ; B # ARABIC LETTER YEH BARREE
06D5 ; B # ARABIC LETTER AE
<dotless noon> ; B # ARABIC LETTER DOTLESS NOON
#
# Sequences that are discouraged in normal use, being indistinguishable from preexisting letters:
#
066E <one dot below> ; X # 0628 # ARABIC LETTER BEH
06D5 <two dots horiz above> ; X # 0629 # ARABIC LETTER TEH MARBUTA
066E <two dots horiz above> ; X # 062A # ARABIC LETTER TEH
066E <three dots up above> ; X # 062B # ARABIC LETTER THEH
062D <one dot below> ; X # 062C # ARABIC LETTER JEEM
062D <one dot above> ; X # 062E # ARABIC LETTER KHAH
062F <one dot above> ; X # 0630 # ARABIC LETTER THAL
0631 <one dot above> ; X # 0632 # ARABIC LETTER ZAIN
0633 <three dots up above> ; X # 0634 # ARABIC LETTER SHEEN
0635 <one dot above> ; X # 0636 # ARABIC LETTER DAD
0637 <one dot above> ; X # 0638 # ARABIC LETTER ZAH
0639 <one dot above> ; X # 063A # ARABIC LETTER GHAIN
06A1 <one dot above> ; X # 0641 # ARABIC LETTER FEH
066F <two dots horiz above> ; X # 0642 # ARABIC LETTER QAF
<dotless noon> <one dot above> ; X # 0646 # ARABIC LETTER NOON
0649 <two dots horiz below> ; X # 064A # ARABIC LETTER YEH # *** Questionable because of use of 064A
in canonical decomposition of 0626.
066E <tah above> ; X # 0679 # ARABIC LETTER TTEH
066E <two dots vert above> ; X # 067A # ARABIC LETTER TTEHEH
066E <two dots vert below> ; X # 067B # ARABIC LETTER BEEH
066E <ring below> <two dots horiz above> ; X # 067C # ARABIC LETTER TEH WITH RING
066E <two dots horiz above> <ring below> ; X # 067C # ARABIC LETTER TEH WITH RING
066E <three dots down above> ; X # 067D # ARABIC LETTER TEH WITH THREE DOTS ABOVE DOWNWARDS
066E <three dots down below> ; X # 067E # ARABIC LETTER PEH
066E <four dots above> ; X # 067F # ARABIC LETTER TEHEH
066E <four dots below> ; X # 0680 # ARABIC LETTER BEHEH
062D 0654 ; X # 0681 # ARABIC LETTER HAH WITH HAMZA ABOVE
062D <two dots vert above> ; X # 0682 # ARABIC LETTER HAH WITH TWO DOTS VERTICAL ABOVE
062D <two dots horiz below> ; X # 0683 # ARABIC LETTER NYEH
062D <two dots vert below> ; X # 0684 # ARABIC LETTER DYEH
062D <three dots up above> ; X # 0685 # ARABIC LETTER HAH WITH THREE DOTS ABOVE
062D <three dots down below> ; X # 0686 # ARABIC LETTER TCHEH
062D <four dots below> ; X # 0687 # ARABIC LETTER TCHEHEH
062F <tah above> ; X # 0688 # ARABIC LETTER DDAL
062F <ring below> ; X # 0689 # ARABIC LETTER DAL WITH RING
062F <one dot below> ; X # 068A # ARABIC LETTER DAL WITH DOT BELOW
062F <one dot below> <tah above> ; X # 068B # ARABIC LETTER DAL WITH DOT BELOW AND SMALL TAH
062F <tah above> <one dot below> ; X # 068B # ARABIC LETTER DAL WITH DOT BELOW AND SMALL TAH
062F <two dots horiz above> ; X # 068C # ARABIC LETTER DAHAL
062F <two dots horiz below> ; X # 068D # ARABIC LETTER DDAHAL
062F <three dots up above> ; X # 068E # ARABIC LETTER DUL
062F <three dots down above> ; X # 068F # ARABIC LETTER DAL WITH THREE DOTS ABOVE DOWNWARDS
062F <four dots above> ; X # 0690 # ARABIC LETTER DAL WITH FOUR DOTS ABOVE
0631 <tah above> ; X # 0691 # ARABIC LETTER RREH
0631 <small v above> ; X # 0692 # ARABIC LETTER REH WITH SMALL V
0631 <ring below> ; X # 0693 # ARABIC LETTER REH WITH RING
0631 <one dot below> ; X # 0694 # ARABIC LETTER REH WITH DOT BELOW
0631 <small v below> ; X # 0695 # ARABIC LETTER REH WITH SMALL V BELOW
0631 <one dot below> <one dot above> ; X # 0696 # ARABIC LETTER REH WITH DOT BELOW AND DOT ABOVE
0631 <one dot above> <one dot below> ; X # 0696 # ARABIC LETTER REH WITH DOT BELOW AND DOT ABOVE
0631 <two dots horiz above> ; X # 0697 # ARABIC LETTER REH WITH TWO DOTS ABOVE
0631 <three dots up above> ; X # 0698 # ARABIC LETTER JEH
0631 <four dots above> ; X # 0699 # ARABIC LETTER REH WITH FOUR DOTS ABOVE
0633 <one dot below> <one dot above> ; X # 069A # ARABIC LETTER SEEN WITH DOT BELOW AND DOT ABOVE
0633 <one dot above> <one dot below> ; X # 069A # ARABIC LETTER SEEN WITH DOT BELOW AND DOT ABOVE
0633 <three dots down below> ; X # 069B # ARABIC LETTER SEEN WITH THREE DOTS BELOW

```

0633 <three dots down below> <three dots up above> ; X # 069C # ARABIC LETTER SEEN WITH THREE DOTS BELOW AND THREE DOTS ABOVE

0633 <three dots up above> <three dots down below> ; X # 069C # ARABIC LETTER SEEN WITH THREE DOTS BELOW AND THREE DOTS ABOVE

0635 <two dots horiz below> ; X # 069D # ARABIC LETTER SAD WITH TWO DOTS BELOW

0635 <three dots up above> ; X # 069E # ARABIC LETTER SAD WITH THREE DOTS ABOVE

0637 <three dots up above> ; X # 069F # ARABIC LETTER TAH WITH THREE DOTS ABOVE

0639 <three dots up above> ; X # 06A0 # ARABIC LETTER AIN WITH THREE DOTS ABOVE

06A1 <one dot below> ; X # 06A2 # ARABIC LETTER FEH WITH DOT MOVED BELOW

06A1 <one dot below> <one dot above> ; X # 06A3 # ARABIC LETTER FEH WITH DOT BELOW

06A1 <one dot above> <one dot below> ; X # 06A3 # ARABIC LETTER FEH WITH DOT BELOW

06A1 <three dots up above> ; X # 06A4 # ARABIC LETTER VEH

06A1 <three dots down below> ; X # 06A5 # ARABIC LETTER FEH WITH THREE DOTS BELOW

06A1 <four dots above> ; X # 06A6 # ARABIC LETTER PEHEH

066F <one dot above> ; X # 06A7 # ARABIC LETTER QAF WITH DOT ABOVE

066F <three dots up above> ; X # 06A8 # ARABIC LETTER QAF WITH THREE DOTS ABOVE

06A9 <ring below> ; X # 06AB # ARABIC LETTER KAF WITH RING

0643 <one dot above> ; X # 06AC # ARABIC LETTER KAF WITH DOT ABOVE

0643 <three dots up above> ; X # 06AD # ARABIC LETTER NG

0643 <three dots down below> ; X # 06AE # ARABIC LETTER KAF WITH THREE DOTS BELOW

06AF <ring below> ; X # 06B0 # ARABIC LETTER GAF WITH RING

06AF <two dots horiz above> ; X # 06B1 # ARABIC LETTER NGOEH

06AF <two dots horiz below> ; X # 06B2 # ARABIC LETTER GAF WITH TWO DOTS BELOW

06AF <two dots vert below> ; X # 06B3 # ARABIC LETTER GUEH

06AF <three dots up above> ; X # 06B4 # ARABIC LETTER GAF WITH THREE DOTS ABOVE

0644 <small v above> ; X # 06B5 # ARABIC LETTER LAM WITH SMALL V

0644 <one dot above> ; X # 06B6 # ARABIC LETTER LAM WITH DOT ABOVE

0644 <three dots up above> ; X # 06B7 # ARABIC LETTER LAM WITH THREE DOTS ABOVE

0644 <three dots down below> ; X # 06B8 # ARABIC LETTER LAM WITH THREE DOTS BELOW

<dotless noon> <one dot below> <one dot above> ; X # 06B9 # ARABIC LETTER NOON WITH DOT BELOW

<dotless noon> <one dot above> <one dot below> ; X # 06B9 # ARABIC LETTER NOON WITH DOT BELOW

<dotless noon> <tah above> ; X # 06BB # ARABIC LETTER RNOON

<dotless noon> <ring below> <one dot above> ; X # 06BC # ARABIC LETTER NOON WITH RING

<dotless noon> <one dot above> <ring below> ; X # 06BC # ARABIC LETTER NOON WITH RING

<dotless noon> <three dots up above> ; X # 06BD # ARABIC LETTER NOON WITH THREE DOTS ABOVE # ***

Excluded because of non-standard linking behavior; dots go below in initial and medial forms.

062D <three dots down below> <one dot above> ; X # 06BF # ARABIC LETTER TCHEH WITH DOT ABOVE

062D <one dot above> <three dots down below> ; X # 06BF # ARABIC LETTER TCHEH WITH DOT ABOVE

06C1 <two dots horiz above> ; X # 06C3 # ARABIC LETTER TEH MARBUTA GOAL

0648 <ring above> ; X # 06C4 # ARABIC LETTER WAW WITH RING

0648 <bar through> ; X # 06C5 # ARABIC LETTER KIRGHIZ OE

0648 <small v above> ; X # 06C6 # ARABIC LETTER OE

0648 <inverted v above> ; X # 06C9 # ARABIC LETTER KIRGHIZ YU

0648 <two dots horiz above> ; X # 06CA # ARABIC LETTER WAW WITH TWO DOTS ABOVE

0648 <three dots up above> ; X # 06CB # ARABIC LETTER VE

06CC <small v above> ; X # 06CE # ARABIC LETTER YEH WITH SMALL V

0648 <one dot above> ; X # 06CF # ARABIC LETTER WAW WITH DOT ABOVE

0649 <two dots vert below> ; X # 06D0 # ARABIC LETTER E

0649 <three dots down below> ; X # 06D1 # ARABIC LETTER YEH WITH THREE DOTS BELOW

062F <inverted v above> ; X # 06EE # ARABIC LETTER DAL WITH INVERTED V

0631 <inverted v above> ; X # 06EF # ARABIC LETTER REH WITH INVERTED V

0633 <one dot below> <three dots up above> ; X # 06FA # ARABIC LETTER SHEEN WITH DOT BELOW

0633 <three dots up above> <one dot below> ; X # 06FA # ARABIC LETTER SHEEN WITH DOT BELOW

0635 <one dot below> <one dot above> ; X # 06FB # ARABIC LETTER DAD WITH DOT BELOW

0635 <one dot above> <one dot below> ; X # 06FB # ARABIC LETTER DAD WITH DOT BELOW

0639 <one dot below> <one dot above> ; X # 06FC # ARABIC LETTER GHAIN WITH DOT BELOW

0639 <one dot above> <one dot below> ; X # 06FC # ARABIC LETTER GHAIN WITH DOT BELOW

06BE <inverted v above> ; X # 06FF # ARABIC LETTER HEH WITH INVERTED V

Appendix B: Default collation keys

Suggested collation key values for the proposed Arabic modifier marks are given in the file *ArabicModifierKeys.txt*. The values used here are based on those found in version 3.1.1 of the standard *allkeys.txt* file.

The draft data file listed here uses abbreviated names to refer to proposed characters for which Unicode codepoints are not yet available.

Listing of data file *ArabicModifierKeys.txt*

```
# Suggested default key weights for Arabic modifier marks, based on version 3.1.1 of allkeys.txt
#
# Individual characters; 4th field to be Unicode codepoint
#
<single dot above> ; [.0000.0200.0002.XXXX] # ARABIC MODIFIER SINGLE DOT ABOVE
<two dots horiz above> ; [.0000.0201.0002.XXXX] # ARABIC MODIFIER TWO DOTS HORIZONTALLY ABOVE
<two dots vert above> ; [.0000.0202.0002.XXXX] # ARABIC MODIFIER TWO DOTS VERTICALLY ABOVE
<three dots up above> ; [.0000.0203.0002.XXXX] # ARABIC MODIFIER THREE DOTS UPWARDS ABOVE
<three dots down above> ; [.0000.0204.0002.XXXX] # ARABIC MODIFIER THREE DOTS DOWNWARDS ABOVE
<four dots above> ; [.0000.0205.0002.XXXX] # ARABIC MODIFIER FOUR DOTS ABOVE
<tah above> ; [.0000.0210.0002.XXXX] # ARABIC MODIFIER TAH ABOVE
<small v above> ; [.0000.0211.0002.XXXX] # ARABIC MODIFIER SMALL V ABOVE
<inverted v above> ; [.0000.0212.0002.XXXX] # ARABIC MODIFIER INVERTED V ABOVE
<ring above> ; [.0000.0213.0002.XXXX] # ARABIC MODIFIER RING ABOVE
<single dot below> ; [.0000.0220.0002.XXXX] # ARABIC MODIFIER SINGLE DOT BELOW
<two dots horiz below> ; [.0000.0221.0002.XXXX] # ARABIC MODIFIER TWO DOTS HORIZONTALLY BELOW
<two dots vert below> ; [.0000.0222.0002.XXXX] # ARABIC MODIFIER TWO DOTS VERTICALLY BELOW
<three dots up below> ; [.0000.0223.0002.XXXX] # ARABIC MODIFIER THREE DOTS UPWARDS BELOW
<three dots down below> ; [.0000.0224.0002.XXXX] # ARABIC MODIFIER THREE DOTS DOWNWARDS BELOW
<three dots horiz below> ; [.0000.0225.0002.XXXX] # ARABIC MODIFIER THREE DOTS HORIZONTALLY BELOW
<four dots below> ; [.0000.0226.0002.XXXX] # ARABIC MODIFIER FOUR DOTS BELOW
<tah below> ; [.0000.0230.0002.XXXX] # ARABIC MODIFIER TAH BELOW
<small v below> ; [.0000.0231.0002.XXXX] # ARABIC MODIFIER SMALL V BELOW
<inverted v below> ; [.0000.0232.0002.XXXX] # ARABIC MODIFIER INVERTED V BELOW
<ring below> ; [.0000.0233.0002.XXXX] # ARABIC MODIFIER RING BELOW
<bar through> ; [.0000.0240.0002.XXXX] # ARABIC MODIFIER BAR THROUGH
#
# DOTLESS NOON goes in the main sequence right after ARABIC LETTER NOON.
# All Level 1 weights from 0F34 (ARABIC LETTER NOON GHUNNA) upwards will then be increased by 1.
#
<dotless noon> ; [.0F34.0020.0002.XXXX] # ARABIC LETTER DOTLESS NOON
```
