**L2/02-061**

# Bidi committee consensus on Arabic additions from L2/01-425

*Jonathan Kew, SIL International*
*January 29, 2002*

There were 13 characters proposed in L2/01-425, and there seemed to be general agreement among the Bidi committee, after some initial discussion, that they were all valid candidates for encoding. An updated version of the summary from L2/01-425 is presented here, incorporating suggestions from a number of the Bidi experts.

| *Representative glyph* | *Suggested USV* | *Character name* | *General category* | *Combining class* | *Bidi category* |
|---|---|---|---|---|---|
|  | U+0600 | ARABIC NUMBER SIGN | Cf | 0 | AL |
|  | U+0601 | ARABIC YEAR SIGN | Cf | 0 | AL |
|  | U+0602 | ARABIC FOOTNOTE MARKER | Cf | 0 | AL |
|  | U+060D | ARABIC POETIC VERSE SIGN | Po | 0 | ON |
|  | U+060E | ARABIC DATE SEPARATOR | Po | 0 | AL |
|  | U+0610 | ARABIC SIGN SALLALLAHOU ALAYHE WASALLAM | Mn | 250 | NSM |
|  | U+0611 | ARABIC SIGN ALAYHE ASSALAM | Mn | 250 | NSM |
|  | U+0612 | ARABIC SIGN RAHMATULLAH ALAYHE | Mn | 250 | NSM |
|  | U+0613 | ARABIC SIGN RADI ALLAHOU ANHU | Mn | 250 | NSM |
|  | U+0614 | ARABIC SIGN NOM DE PLUME | Mn | 250 | NSM |
|  | U+0656 | ARABIC SUBSCRIPT ALEF | Mn | 37 | NSM |
|  | U+0657 | ARABIC TURNED DAMMA | Mn | 38 | NSM |
|  | U+0658 | ARABIC NASALIZATION MARK | Mn | 7 | NSM |

We have left the character names and combining classes largely unchanged from the original proposal, except for a change to improve the consistency of transliteration. There was some discussion of adding "URDU" in the character names, but given that (a) the marks are used in other languages besides Urdu, and (b) no additional language name is used in the names most other "extended" Arabic-script characters, we do not see that this is necessary or helpful.

While the ARABIC POETIC VERSE SIGN and ARABIC FOOTNOTE MARKER are similar in appearance, there are many differences in function and behavior, and therefore two distinct characters should be encoded:

ARABIC POETIC VERSE SIGN: This is an ordinary spacing symbol that does not combine with anything. It functions basically like a bullet, set off somewhat in the margin to indicate one or more lines of Arabic verse. (See, for example, Figure 18 in L2/01-426.)

ARABIC FOOTNOTE MARKER: This is an enclosing symbol that encloses one or more Arabic digits to indicate a footnote. The marker plus digits are placed in text at the point of annotation, to number a footnote. They occur again at the bottom or margin of text to indicate the start of the corresponding numbered footnote. The ARABIC FOOTNOTE MARKER precedes (in logical order) the digits it encloses (usually just one or two), and the length of the bottom stroke is adjusted to subtend the sequence of digits it encloses. (Figure 30 in L2/01-426 shows many examples.)

In the proposed classes, we have attempted to follow existing patterns of combining class assignments as far as possible. In particular:

- Vowel marks are assigned to "fixed position" classes following the existing Arabic vowel marks;
- The ARABIC NASALIZATION MARK is considered equivalent to a "nukta", as it is a modifier that binds tightly to the underlying letter;
- The honorifics (U+0610..0614) are assigned a new class 250, placing them further from the base character than any other marks, as they are really "word-level" rather than "character-level" marks.

In addition to the proposed new characters, we recommend that an annotation be added to the Standard for U+06E1: *alternate name ARABIC JAZM; alternate form of U+0652 ARABIC SUKUN, but used for distinct purposes in some contexts*. We also recommend that U+06DD ARABIC END OF AYAH be changed to use a "Syriac Abbreviation Mark" model, along with the new characters proposed for U+0600..0602 (see below), as an enclosing mark behavior is cumbersome to use for a character that routinely applies to digit sequences, not just single base characters.

The use of category Cf for the digit-related marks, including the change to END OF AYAH, follows Ken Whistler's email message of 8-Jan-02 and the expressed preference of the UTC: *[89-C27] Consensus: For the ARABIC END OF AYAH the UTC prefers the head model rather than the tail model, like the Syriac abbreviation mark. [L2/01-428].*

Three possible approaches to these marks have been considered:

A) No change (for END OF AYAH; other marks would follow this example). If multiple digits are to be enclosed, a U+034F COMBINING GRAPHEME JOINER must be inserted between the digits.

B) Change to use the SAM model. The ARABIC END OF AYAH would apply to characters after it, up some well-defined break point. This would mean that the ARABIC END OF AYAH must be moved before the sequence of characters that it is to apply to, and would enclose the entire word.

C) Change to use an inverse SAM model. The ARABIC END OF AYAH would apply to characters before it, back to some well-defined break point.

Option A requires the insertion of CGJ in the common case of enclosing a multi-digit number. This is a cumbersome and non-intuitive requirement for users, and a likely outcome is that implementers will ignore it and implement something closer to what users would expect, such as enclosing a preceding sequence of up to three digits. (See Uniscribe.) Conformant implementations will be unfriendly to users; user-friendly implementations will be non-conformant and possibly inconsistent with each other.

Either B or C could, with suitable definition of the "well-defined break point", provide a more user-friendly encoding convention, without the requirement to add invisible control characters (CGJ or others) in normal usage. In the case of the YEAR, NUMBER, and FOOTNOTE marks, at least, it is linguistically more logical to encode the mark preceding the number; this favors option B rather than C. In addition, it seems preferable to avoid creating yet another model, where the existing SAM model is in any case preferred for semantic reasons.

The only disadvantage of option B would seem to be the requirement to move END OF AYAH from after to before the number in any existing data. However, given that there is not currently any conformant way to encode an END OF AYAH that encloses multiple digits, and few (if any) implementations in widespread use, this does not constitute a serious barrier to adopting this model.

We have proposed AL rather than BN as the bidi type of these characters (note that the SAM, in contrast, is currently BN), as it seems that they should have a strong R directionality in order to ensure consistent behavior during layout. Implementation will be more difficult and confusing if it is not guaranteed that the mark occurs to the right of the number after bidi reordering, and citing such a form within Latin text, for example, would probably require the addition of RLM to control the directionality of the mark.

These four "number adornments" of category Cf (the three new characters and the END OF AYAH) are defined to apply to the following grapheme cluster or digit sequence. This proposal is based on the fact that the most common case for all of these enclosing marks will be to enclose a sequence of digits, possibly followed by a single character (particularly in the case of the YEAR MARK). On occasion the digit sequence may include things such as thousands, decimal separators or dashes. One can conceive of even more complex needs that we don't want to prohibit, but at the same time we're willing to require more cumbersome encoding for these.

It seems natural to suggest that these marks would apply to the following "word" (like the SAM), but this is not a satisfactory definition, given the lack of a clear definition of a word boundary for this purpose. We considered an approach based on line-break properties, but this falls foul of the potential need to control the scope of the enclosing mark separately from controlling line breaks. We also recognize that these marks are not normally applied to arbitrary text, so a more limited definition than "word" is appropriate.

The proposal, then, is that the scope of the mark is defined to be the immediately following grapheme cluster, with the addition that if the grapheme cluster ends with a digit, it is automatically extended to include all immediately succeeding digits, just as if there were a COMBINING GRAPHEME JOINER between each such pair of digits. ("Digit" here means a character with general category Nd.) More formally, let us define a "simple grapheme" (SG) as a grapheme that does not include CGJ characters. Then we can say that the scope of the mark is the following EnclosableCluster, where:

EnclosableGrapheme = (Digit Digit*) | SG
EnclosableCluster = EnclosableGrapheme (CGJ EnclosableGrapheme)*

Using this proposal, a mark applied to a sequence of digits requires no CGJs, while a sequence of digits followed by a single letter, such as needed for a year + era use, would require a single CGJ.

We recognize that some software may wish to implement a subset of these capabilities. The following are suggested as minimal capabilities for rendering software that supports these characters:

- for YEAR: at least 4 digits plus, if present, the following CGJ and single Arabic character
- for END OF AYAH and FOOTNOTE: at least three digits
- for NUMBER: at least four digits

Implementations that follow these guidelines should be adequate for most normal usage. However, the Standard should not preclude an implementation supporting additional combinations that might be needed by specialized users, such as the use of Arabic letters to enumerate footnotes, or internal punctuation (thousands and decimal separators) within the marked number.

In UCD data file format, the entries for the new characters would be:

```
0600;ARABIC NUMBER SIGN;Cf;0;AL;;;;;N;;;;;
0601;ARABIC YEAR SIGN;Cf;0;AL;;;;;N;;;;;
0602;ARABIC FOOTNOTE MARKER;Cf;0;AL;;;;;N;;;;;

060D;ARABIC POETIC VERSE SIGN;Po;0;ON;;;;;N;;;;;
060E;ARABIC DATE SEPARATOR;Po;0;AL;;;;;N;;;;;

0610;ARABIC SIGN SALLALLAHOU ALAYHE WASALLAM;Mn;250;NSM;;;;;N;;;;;
0611;ARABIC SIGN ALAYHE ASSALAM;Mn;250;NSM;;;;;N;;;;;
0612;ARABIC SIGN RAHMATULLAH ALAYHE;Mn;250;NSM;;;;;N;;;;;
0613;ARABIC SIGN RADI ALLAHOU ANHU;Mn;250;NSM;;;;;N;;;;;
0614;ARABIC SIGN NOM DE PLUME;Mn;250;NSM;;;;;N;;;;;

0656;ARABIC SUBSCRIPT ALEF;Mn;37;NSM;;;;;N;;;;;
0657;ARABIC TURNED DAMMA;Mn;38;NSM;;;;;N;;;;;
0658;ARABIC NASALIZATION MARK;Mn;7;NSM;;;;;N;;;;;
```

In addition, we propose that:

```
06DD;ARABIC END OF AYAH;Me;0;NSM;;;;;N;;;;;
```

be changed to:

```
06DD;ARABIC END OF AYAH;Cf;0;AL;;;;;N;;;;;
```

Finally, it should be noted that during discussion of this proposal, several additional characters have been mentioned, and we expect that once adequate documentation is put forward, these will be proposed as further additions. This proposal should not be interpreted as aiming to “complete” the set of Arabic-script characters required.