L2/01-428

Request for clarification regarding U+06DD ARABIC END OF AYAH and other Arabic enclosing marks

Date: November 1, 2001
Author: Jonathan Kew, SIL International
Address: Horsleys Green
High Wycombe
Bucks HP14 3XL
England
Tel: +44 1494 682306
Email: jonathan_kew@sil.org

The character U+06DD ARABIC END OF AYAH is defined as an *enclosing mark* (Me). Normally, any combining mark combines with the previous base character; thus, one would expect END OF AYAH to enclose a single base character such as a digit.

In this case, however, the intended use of the mark is to enclose ayah (verse) numbers in the Qur'an, which may have up to three digits. It seems logical, then, that the END OF AYAH character should enclose the preceding digit sequence (not just a single base character). This is the behavior being implemented, for example, in Uniscribe.

In order to encourage consistent implementation and encoding of data, it would be helpful to have a statement clarifying the correct combining behavior of END OF AYAH:

- Does this character enclose digit sequences or only single base characters? If the latter, how would a multi-digit ayah number be encoded—with grapheme joiners? This seems highly undesirable from a user's point of view.
- How many digits may be enclosed? Should there be a limit of three digits (the most found in Qur'anic ayah numbers), or should it (in principle, at least) enclose an arbitrarily long digit sequence?
- Exactly which characters count as members of a digit sequence? Arabic digits? European digits? Digits of other scripts? Can any non-digit characters such as number separator or terminator characters be considered part of the sequence?

Similar considerations will apply to the proposed characters ARABIC YEAR SIGN, ARABIC NUMBER SIGN, and ARABIC FOOTNOTE MARKER. In all these cases, the natural expectation would be that the mark interacts with a digit sequence, not just a single base character. In the case of the year sign, users may want the character representing the "era" (Arabic-script equivalents of AD, BC, AH) to be "enclosed" by the mark, as well as the actual year number. For the number sign, thousands separators (and perhaps a decimal separator) might be part of the number. And for the footnote marker, it is possible that users might want non-numeric footnote references.

In each of these cases, it would be helpful to have a statement giving the intended behavior. Assuming these characters are to be exceptions to the rule of combining with the one preceding base character, the exact nature of their exceptional behavior must be made clear.