International

NRSI
Wycliffe Centre
Horsleys Green
High Wycombe
Bucks HP14 3XL
England

Tel:
+44 (1494) 682306

Fax:
+44 (1494) 682300

Email:
jonathan_kew@sil.org

Dr. Attash Durrani
Center of Excellence for Urdu Informatics
National Language Authority
Pakistan

L2/06-240

cc: L2 document register (l2doc@unicode.org)

July 19, 2006

Dear Dr. Durrani,

I am writing at the request of the Unicode Technical Committee with comments in response to your document L2/06-039 *Preliminary Proposal to add Nuqta Characters to Arabic Block*.

This document presents a concept that has been considered in the past: that Arabic-script letters could be encoded as combinations of “skeletal” letterforms such as ٮ and ح with dots (or sometimes other marks) positioned above or below, such as ◌̇, ◌ٜ, or ◌ؕ. It is understood that this would be consistent with the inherent structure and historical development of the Arabic script, and would be helpful where regional languages are adopting new ways of writing using combinations of basic letter plus dots that have not previously been used. This approach can be referred to as a “generative” encoding of Arabic.

However, the established encoding practice in Unicode/ISO 10646 is that Arabic letters are encoded individually, as non-decomposable units. This practice predates the Unicode standard, which simply followed previous standards (particularly ISO 8859-6) in creating an initial repertoire of Arabic characters. The original repertoire (for Arabic, Persian, Urdu, etc.) has been supplemented with many new characters, as information becomes available.

Many experts now agree that it would have been preferable to use a generative model for Arabic; but at this time, it is not possible to re-design the encoding model used in the Unicode standard. There are a number of reasons why this is so.

Both practical considerations (the amount of existing data and software, all based on the current model) and published stability policies make it impossible to replace the existing Arabic letters with a *base* + *nuqta* encoding. Existing characters cannot under any circumstances be removed from the Standard. (See “Encoding Stability” in the Unicode Stability Policy, http://www.unicode.org/standard/stability_policy.html.)

Therefore, if a set of combining *nuqta* characters were added, there would be two ways to encode most Arabic-script letters: the old individual characters, or the new *base* + *nuqta* sequences. This is not good for users, as it would lead to confusion and ambiguity, and to problems such as spoofing.

Where Unicode does, in fact, have two ways to encode a given piece of text (accented Latin characters being a primary example), the concepts of canonical equivalence and normalization are there to solve most of the problems this could cause. Thus, Unicode specifies that U+0101 *ā* has a canonical decomposition to <U+0061 *a*, U+0304 ◌̄>. This means that all systems can treat these representations as equivalent, and convert between them as needed.

It is natural to suggest, then, that the existing Arabic letters should be given canonical decompositions to *base* + *nuqta* sequences, just as the accented Latin letters have. Thus, U+062A ت would have a decomposition to <U+066E ٮ, U+XXXX ◌̈>, etc. However, stability considerations do not allow existing characters to be given decompositions where they do not already have them. (See "Normalization Stability" in the Policy.) Any change that would disrupt normalization stability could have bad effects on other standards and protocols that depend on Unicode, and could lead to major interoperability problems between systems. Therefore, the "obvious" solution is not available to us.

Several years ago, in fact, I co-authored a proposal, with input from others on the Committee, for a set of characters similar to the *nuqta* characters you have requested; see document L2/03-154. At that time, a proposal was also presented for extending the normalization process (L2/03-171) so that it would be possible to define a canonical equivalence between the existing Arabic letters and their *base* + *nuqta* representations, without destabilizing the normalization forms. These proposals were considered by the UTC in June 2003. After considerable discussion, the final decision was that even though such an encoding model would have real benefits, the costs of making such a change at this stage would be too great. I think this related both to technical (implementation) requirements and to the confusion that introducing a parallel encoding model could cause. Ultimately, motion 95-M1 *In principle the UTC is in favor of encoding additional diacritic combining characters for productive extension of the Arabic script* failed to pass, and I do not think the committee is likely to reconsider this question now.

This is, then, the clear policy in Unicode: Arabic letters are encoded as indivisible units, including their *nuqta* combinations, and are not composed dynamically. As a result, new letters continue to be added to the Standard, as documentation is presented; see the Arabic Supplement block at U+075x.

In passing, note that there are some cases where an Arabic-script letter is visually similar to a base letter plus an existing diacritic such as U+0654 *hamza above* or U+0615 *small high tah*. In cases like U+0623 أ or U+0626 ئ, where the meaning is indeed a letter with hamza added, these characters have canonical decompositions. But in the case of U+0681 ځ, U+076C ݬ or U+0759 ݙ where the *hamza* or *tah* shape is used as part of a new letter, not as a separate letter or mark in its own right, there is no decomposition; these characters are encoded as distinct letters.

Your document also questioned the purpose of encoding the dotless forms U+066E ٮ and U+066F ٯ, when there are no combining *nuqta* marks that can be added to form complete letters. These were encoded to complete the set of dotless skeletal letters so that archaic undotted Arabic text can be represented in Unicode; they are not expected to function as base characters for the construction of modern dotted letters.

It is important to remember that the use of a "precomposed" encoding model for Arabic letters in Unicode does not prevent font developers using decomposed forms and dynamic composition technology to actually render the text. In complex styles such as Nastaliq, a typical OpenType font implementation may begin with a glyph decomposition step, mapping each letter to a base form and one or more combining mark glyphs; then contextual replacement and positioning rules can be applied to these. But this happens entirely within the font, using font-specific glyph codes, and is not related to the Unicode encoding.

To conclude, then, the combining *nuqta* characters are not regarded as valid candidates for addition to the Standard. That would be a very reasonable model for encoding the Arabic script, but it is not the model used in Unicode.

In the light of this, I would invite you to submit proposals for any specific characters required that are not currently available in Unicode, whether these are for regional languages or archaic documents. In particular, I think the *reh with four dots below* shown in Figure 2 of your document is not yet encoded, and there may be others as well.

In the case of the double lines above or below letters, used with *seen* and *hah* base shapes, I think that these may simply represent alternate forms, based on hand-writing, of a four-dot combination. In Shina, at least, I am familiar with the use of both *seen with four dots above* and *seen with two lines above* to represent a retroflexed ṣ, and I suspect that the same may be true for *ch* written as *hah with four dots below*. It would be debatable, therefore, whether the two-line combination should be encoded separately, or merely regarded as a glyph alternate for four dots.

Finally, if there are additional diacritic marks that are used as diacritics, similarly to the Arabic vowel marks, Qur'anic annotation marks, etc., these can of course be considered for addition to Unicode. But if they are used to create new Arabic-script consonants, along the pattern of the *nuqtas* and *small tah*, then such letters should be proposed as individual characters.

Regards,

Jonathan Kew