L2/03-166

Proposal for encoding the combining diacritic ARABIC WASLA

Miikka-Markus Alhonen

May 28, 2003

A. Administrative

1. Title
Proposal for encoding the combining diacritic ARABIC WASLA.
2. Requester's name
Miikka-Markus Alhonen.
3. Requester type
Individual contribution.
4. Submission date
2003-05-28.
5. Requester's reference
6a. Completion
This is a complete proposal.
6b. More information to be provided?
No.

B. Technical – General

1a. New script? Name?
No.
1b. Addition of characters to existing block? Name?
Yes. Arabic.
2. Number of characters
1
3. Proposed category
Category A.
4. Proposed level of implementation and rationale
Level 3, a combining character.
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?
Thomas Milo, DecoType.
6b. Font currently available?
Yes.
6c. Font format?
True Type.
7a. Are references (to other character sets, dictionaries, descriptive texts, etc.) provided?
Yes.
7b. Are published examples (such as newspapers, magazines, or other sources) of use of proposed characters attached?
Yes.
8. Does the proposal address other aspects of character data processing?
No.

C. Technical – Justification

1. Has this proposal for addition of character(s) been submitted before?
No.
2. Contact with the user community?
Yes. Scholars at Helsinki University, experts in the Syriac script, and members of the Arabeyes community (`http://www.arabeyes.org/`).
3. Information on the user community?
Speakers of the Arabic language, scholars.
4a. The context of use for the proposed characters?
To write the Arabic language in Arabic or Syriac script.
4b. Reference
5a. Proposed characters in current use?
Yes.
5b. Where?
Worldwide.
6a. Characters should be encoded entirely in BMP?
Yes. Position U+0659 is proposed.
5b. Rationale
To keep it together with the rest of the Arabic script, contemporary use.
6. Should characters be kept in a continuous range?
N/A
7a. Can the characters be considered a presentation form of an existing character or character sequence?
No.
7b. Where?

7c. Reference
8a. Can any of the characters be considered to be similar (in appearance or function) to an existing character?
No.
8b. Where?
8c. Reference
9a. Combining characters or use of composite sequences included?
Yes.
9b. List of composite sequences and their corresponding glyph images provided?
No.
10. Characters with any special properties such as control function, etc. included?
No.

D. SC 2/WG 2 Administrative (To be completed by SC 2/WG 2)

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

E. Proposal

This is a proposal for encoding the diacritical mark ARABIC WASLA in the Unicode standard. The character is mostly used in the combination U+0627 ARABIC LETTER ALEF + ARABIC WASLA, which can be encoded as U+0671 ARABIC LETTER ALEF WASLA at present. This proposal discusses the benefits of assigning a new code point for the diacritic itself, thus allowing for a full range of combinations to be represented in Unicode.

The Arabic *waṣla* وصلة—also known as *ṣila* صلة or *hamzatu-l-waṣl* همزة الوصل—is a combining diacritic, which is used together with other diacritical marks to indicate the exact pronunciation of an Arabic text. These diacritics include the short vowels *fatḥa*, *ḍamma* and *kasra*, as well as the sign for vowellessness (*sukūn*), the glottal stop (*hamza*) and various others. All of the regularly used diacritics are already encoded in the UCS with the sole exception of *waṣla*.

The use of *waṣla* is in some ways very different from the other diacritics. First of all, it does not normally occur with any other letter than U+0627 *ʾalef*. It is written above *ʾalef* to indicate the absence of a word-initial glottal stop, marked with *hamza* (U+0654 or U+0655), and the vowel following it. This

omission is entirely predictable, as it occurs *always* in the middle of a sentence if the *ʾalef* belongs to the definite article, or if the word is a certain kind of an imperative or a perfect tense verb, a verbal noun or one of the very few exceptional common nouns. For this reason, its use is *never* obligatory, since anyone with minimal knowledge of the Arabic language will know where to place the mark if needed. This demonstrates the diacritical nature of the mark, since it is not to be considered an integral part of a single character called *ʾalef waṣla* but merely an optional sign for the letter *ʾalef*, which can be written down on some occasions or left out at will. See samples 1 and 2.

Because of the full predictability of the placement of the diacritic, *waṣla* is actually found quite rarely in normal Arabic texts. The main exceptions to this practice are the Bible and the Qur'ān, which are often printed "fully vocalised", i.e. including all the optional vowel signs as well as the rarely marked *waṣla*. This is so due to the need of preserving the sacred religious texts unchanged, as well as due to the fact that not all readers of these books speak Arabic as their native language. Correct pronunciation is thus ensured by marking down everything that a native Arabic speaker would consider obvious in any case. For Qur'ānic and Biblical use, however, the code point U+0671 ARABIC LETTER ALEF WASLA is sufficient, since in normal Arabic text, the diacritic does not occur without *ʾalef*.

The other major usage of *waṣla* can be found in the field of language teaching. Textbooks on Arabic language or the Arabic script in general—the Urdu and the Persian languages, for instance, use some Arabic expressions where *waṣla* is present—often refer to the diacritic by itself or, for instance, on top of U+0640 ARABIC TATWEEL. This behaviour cannot be achieved in any way with the current character repertoire of Unicode or other character set standards. For this reason, some authors have been obliged to draw the diacritic by hand. See samples 3 and 4.

Another possible benefit of assigning a separate codepoint for *waṣla*, is for the so-called *Garshuni* texts, i.e. texts written in the Arabic language but in the Syriac script. As stated in the Unicode Standard, Version 3.0, section 8.3, page 199, *Garshuni employs the Arabic set of vowels and overstrike marks.* Included in these "overstrike marks" is also the ARABIC WASLA. It could well occur combined to U+0710 SYRIAC LETTER ALAPH, although no manuscript samples of this have been found in preparing the original Syriac proposal (N1718) or this document. This is, however, most likely due to the rareness of the diacritic in general than due to the theoretical impossibility of the combination. In fact, I have been informed that the authors of the Syriac proposal agree that the diacritic should be encoded separately, but it was then omitted by mistake. Also some computer implementations have been prepared to handle this case, too. See samples 5 and 6, which are extracted from the manual for the Syriac TEX system *Sabra*. For the use of Garshuni texts, three other Arabic diacritics, namely U+0653 ARABIC MADDA ABOVE, U+0654 ARABIC HAMZA ABOVE and U+0655 ARABIC HAMZA BELOW, were similarly assigned separate code points in Unicode 3.0.

As for the name of the diacritic, it is called *waṣla* ('connexion') by most

orientalists in Western countries. This term is less common in Arabic grammars, where a longer form *hamzatu-l-waṣl* ('a connective *hamza*'), is often used. In Persian and Urdu grammars this is shortened to *waṣl*. Still another form, *ṣila*, occurs in the references of this document, but this name apparently has no contemporary use. For consistency with the name of U+0671 ARABIC LETTER ALEF WASLA, the name ARABIC WASLA is proposed.

For all practical purposes—such as sorting, searching or transliteration—U+0671 ARABIC LETTER ALEF WASLA should be treated as equivalent to U+0627 ARABIC LETTER ALEF + ARABIC WASLA. Due to the Unicode stability policies, however, U+0671 can not be allocated a new canonical decomposition, and it is left for the UTC to decide on how to handle this and other similar cases of misfortunately unassigned decompositions (viz. *Proposal to encode productive Arabic-script modifier marks* by Jonathan Kew, Kamal Mansour and Mark Davis).

Unicode Character Properties

Combining character, category "Mn", bidi category "NSM" (non-spacing mark)

0659 ARABIC WASLA

References

Elementary Modern Standard Arabic. Part 1: Arabic Pronunciation and Writing; Arabic Grammar and Vocabulary, Lessons 1–30. Edited by Abboud, Peter F. & McCarus, Ernest N. Cambridge University Press, Cambridge, 1999.

Fischer, Wolfdietrich. *Grammatik des klassischen Arabisch.* Wiesbaden Harrassowitz, 1972.

Haralambous, Yannis. *Sabra, a Syriac TEX system.* 1996. Manual available at `http://tex.loria.fr/fontes/syriac.ps.gz`

Tikkanen, Bertil. *Urdun kielioppi.* Suomen itämainen seura, Helsinki, 1990.

Annex

2. كَيْفَ ٱلْحالُ means literally "How is the condition?" It is a polite enquiry about health. The response, بِخَيْرٍ , means literally "in (a state of) well-being, or prosperity".

Sample 1. Abboud et al. 1999, page 131. A fully vocalised example sentence with *waṣla*.

كيف الحال ؟	:	نجيب ـ ٣	How are you?
بخير الحمدلله .	:	الاستاذة ـ ٤	Fine, thank you.

Sample 2. Abboud et al. 1999, page 130. The same sentence as in Sample 1 without vocalisation. Note the omission of the *waṣla* sign, too.

den', اخرج *'uḫruǧ*, klass. (*'u*)*ḫruǧ* 'geh hinaus!'. In solchen Fällen wird im Kontext des Klass. Arab. niemals ' gesprochen. Das geschriebene ا ist also leeres Zeichen, was durch ـٰـ (وَصْلَةٌ *waṣla* oder صِلَةٌ *ṣila*) markiert wird: وَٱسْمُهُ *wa-smuhū* 'und sein Name', فَٱنْصَرَفَ *fa-nṣaraf a* 'dann wandte er sich weg', يَا ٱبْنِي *yā bnī* 'o mein Sohn'; im Redebeginn aber: ٱخْرُجْ *'uḫruǧ*.

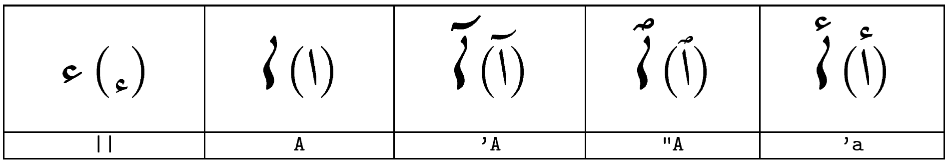
Anm. 1. Das Zeichen ـٰـ ist aus ص, d. h. صِلَةٌ *ṣila* 'Verbindung' entstanden.

Anm. 2. In der arab. Grammatik heißt das *alif*, das im Kontext *waṣla* erhalten muß, أَلِفُ الوَصْلِ *alif al-waṣl*.

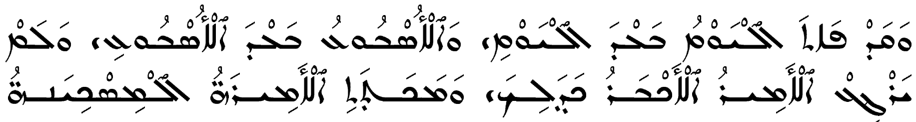
Sample 3. Fischer 1972, page 13. The text explains the origins of the diacritic as well as demonstrates, how it is sometimes placed above U+0640 ARABIC TATWEEL.

9) ٱ **vaṣl** ('liitäntä', ar. **waṣla**) on sidontamerkki, jolla arabian määräävä artikkeli **al-** voidaan liittää edeltävään sanaan, silloin kun se sulautuu edeltävän sanan loppuvokaaliin, esim. امير ٱلمومنين **amīru-l-mūminīn** 'uskovaisten päällikkö'.

Sample 4. Tikkanen 1990, page 22. An example of a hand-drawn *waṣla* in isolation. This quotation is from an Urdu grammar with an Arabic expression used as an example of the diacritic's occurence.



Sample 5. Haralambous 1996, page 14. Table showing the keyboard mapping for SYRIAC LETTER ALAPH + ARABIC WASLA (key sequence "A) in the TEX system *Sabra*.



Sample 6. Haralambous 1996, page 17. Garshuni text with occurences of SYRIAC LETTER ALAPH + ARABIC WASLA.