L2/01-425

Proposal to add Arabic-script honorifics and other marks - WG2 N2483

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

A. Administrative

1. Title	Proposal to add Arabic-script honorifics and other marks
2. Requester's name	SIL International (contacts: Peter Constable, Jonathan Kew)
3. Requester type	Expert contribution
4. Submission date	November 1, 2001
5. Requester's reference	
6a. Completion	This is a complete proposal (when read with Appendix to this proposal, document L2/01-426)
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 in proposal	13
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?	Yes
7b. Are published examples (such as samples from newspapers, magazines, or other sources) of use of proposed characters attached?	Yes
8. Does the proposal address other aspects of character data processing?	Yes, suggested character properties are included

C. Technical — Justification

1. Has this proposal for addition of character(s) been submitted before?	Some of the characters have been submitted previously (cf. L2/00–135), but without comparable documentation
2a. Has contact been made to members of the user community?	Yes
2b. With whom?	During several years in Pakistan, Jonathan Kew worked with Pakistani communities in computerized text editing and publishing. Also see L2/00–135.
3. Information on the user community for the proposed characters is included?	Yes
4. The context of use for the proposed characters	Books published in Urdu, Balochi, and other languages using Arabic script
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 common use
7. Should the proposed characters be kept together in a contiguous range?	No
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 character(s) be considered to be similar (in appearance or function) to an existing character?	No
9b. Rationale for inclusion?	N/A
10. Does the proposal include use of combining characters and/or use of composite sequences?	No
11. Does the proposal contain characters with any special properties?	Yes, most are combining characters

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. Proposal

This proposal presents a number of “marks” used in Arabic script that are not currently included in the UCS repertoire. Most are combining marks, although a couple are non-combining symbols used in conjunction with Arabic script. For ease of understanding, they are presented here in several logical groups.

Figures 1–36, which are scanned pages from published books illustrating the use of these marks, can be found in the appendix to this proposal (supplied as a separate document due to file size).

1. Vowel diacritics

These are vowel marks needed in addition to those already present in the UCS Arabic block.

When Arabic script is adopted as the writing system for a language other than Arabic, the need often arises to represent vowel sounds or distinctions not made in Arabic itself. Conventions including the addition of small dots above and/or below the standard Arabic FATHA, DAMMA, and KASRA signs have been used in some cases. Further investigation is needed to determine whether such conventions are considered “standard” and widely accepted and used in the relevant communities, or whether they are still at an experimental stage.

There are, however, at least two additional vowel marks whose status appears to be well established already, and which should therefore be included in the UCS at this time:

Glyph	*Name*	*Comments*	*Examples*
	ARABIC SUBSCRIPT ALEF	Used to indicate a long /i:/ vowel, or /i/ as contrasted with /e/	See figures 9, 13, 20
	ARABIC TURNED DAMMA	Used to indicate a long /u:/ vowel, or /u/ contrasted with /o/	See figures 13, 16, 21, 23, 26

Classical Arabic has only 3 canonical vowels (/a/, /i/, /u/), while languages such as Urdu and Farsi also have other contrasting vowels such as /o/ and /e/. So, for speakers of these languages, it is imperative to be able to show the difference between /e/ and /i/ (SUBSCRIPT ALEF), and between /o/ and /u/ (TURNED DAMMA). On the other hand, the use of these two diacritics in Arabic is redundant, and serves only to emphasize that the underlying vowel is long.

Considerable email discussion between Kamal Mansour, Thomas Milo and Roozbeh Pournader [17] concerning the document L2/01-304 confirmed the need to encode these characters.

2. SUKUN and JAZM confusion

Closely related to the Arabic vowel marks is the diacritic indicating the absence of a vowel after the base consonant. The ARABIC SUKUN is the normal mark for this purpose, and appears as a “ring” diacritic. In some styles of calligraphy, it is written as a “rotated v” or “hat” shape, with the open side facing either to the left or downwards. In Urdu (and probably some other languages), this form is known as JAZM. I will use the name SUKUN here for the “ring” form, and JAZM for the “open” form, in the hope of avoiding some ambiguity in the discussion. In Naskh-like styles, JAZM is typically open to the left, while in Nastaliq it is typically an inverted “v” or “hat”.

In most circumstances, JAZM can be regarded as a variant of SUKUN, and does not merit separate encoding. However, Tom Milo [18] has pointed out that the Qur’an requires both forms, with the JAZM being the “normal” form used to mark absence of a vowel, and the round SUKUN shape having a different meaning (“ignore this consonant”).

There are currently two characters in the UCS that are relevant here: U+0652 ARABIC SUKUN and U+06E1 ARABIC SMALL HIGH DOTLESS HEAD OF KHAH. These correspond to the Qur’anic forms for “ignore this

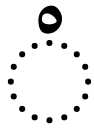
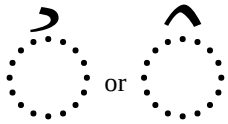
consonant" and "no vowel" respectively. With these two characters, it appears that the Qur'anic requirements are met, although the meaning of SUKUN will not match its conventional Arabic usage.

The encoding of U+06E1 among the "Koranic annotation signs" is somewhat misleading, and its name even more so, but this is now an established fact. The best that can probably be done at this time would be to annotate this character with a note giving the preferred name of ARABIC JAZM and perhaps also a cross-reference to ARABIC SUKUN.

It should be expected that either or possibly both of these characters may vary considerably in form depending on the font style. In a Nastaliq font intended for Urdu, U+06E1 would be rendered as the inverted "v" JAZM shape commonly used in Pakistan. It would probably still be best for U+0652 to be rendered as an Arabic-style SUKUN form. (Note Figure 25, where the Balochi scholar Sayad Hashmi uses both SUKUN and JAZM forms.)

It could be argued, given Tom Milo's explanation of Qur'anic usage, that the Qur'anic JAZM should be considered a glyph variant of U+0652 ARABIC SUKUN, and a new encoded character for ARABIC IGNORED CONSONANT MARK should be added to the UCS. However, given that such a character would look exactly like a standard SUKUN, it seems likely to lead to more rather than less confusion.

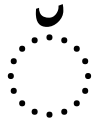
Thus, we propose that explanatory annotations be added for U+06E1, along with a cross-reference to U+0652.

Glyph	*Name*	*Comments*	*Examples*
	(U+0652) ARABIC SUKUN	Marks absence of vowel	See figure 25
	(U+06E1) ARABIC SMALL HIGH DOTLESS HEAD OF KHAH (alternate name: ARABIC JAZM)	Typically used in the Qur'an for SUKUN, but contrasts with the use of the SUKUN form there (for "ignore letter")	See figures 25, 27, 33, 35

3. Nasalization mark

In Urdu and some other languages of Pakistan, there is a need to represent nasalization of a vowel. This is done by writing U+0646 NOON after the vowel. In word final position, the special form U+06BA NOON GHUNNA is used to distinguish nasalization from the consonant /n/. In non-final position, however, no distinction is usually made; the reader generally understands from the context whether nasalization or a consonant is meant. Sometimes, however, the writer wishes to make it unambiguous whether a word-medial NOON represents the consonant /n/ or nasalization of a vowel. To do this, a mark similar to JAZM but oriented as a "cup" shape with the open side upwards is added above the joined (initial or medial) form of NOON (see figures 24, 27, etc.) It is possible that this NASALIZATION MARK was inspired by the SUKUN/JAZM character, but given its distinct form and meaning, it would be inappropriate to unify.

As nasalization is represented in final or isolated position by the character U+06BA NOON GHUNNA (a NOON with no dot), it could be argued that an initial or medial NOON with the NASALIZATION MARK should be regarded as the linking form of NOON GHUNNA, rather than treating the NASALIZATION MARK as an added sign encoded separately. Given that the use of the NASALIZATION MARK is at the discretion of the writer (and is fairly uncommon, at least in Urdu), it seems better to encode it a separate mark that can be added as desired:

Glyph	*Name*	*Comments*	*Examples*
	ARABIC NASALIZATION MARK	Marks NOON as representing vowel nasalization rather than an /n/ consonant	See figures 24, 27, 32, 34

4. Honorifics

These are marks that represent phrases expressing the status of a person. Most have a specifically religious meaning. These marks were proposed by Nelson et al [20], but that document lacked an extensive collection of examples of their use in published books. They are presented here with a number of examples to illustrate their widespread use in the Arabic-script world.

As these marks are in effect combining characters at the word level (rather than being associated with a single base character), they present something of a spelling and canonical ordering problem. It generally seems most logical to add them at the end of the relevant name, but a writer may not always choose to do this. Depending on the letter shapes present in the name and the calligraphic style in use, the writer may add the honorific mark to a letter somewhere in the middle of the name. (Examination of the examples given will show a wide variation in the placement of the mark, suggesting that it could be placed after any of the base characters in the text.) It would be helpful to have a normalization algorithm that could move such "word-level" combining characters to the end of the word, but the present algorithm will not do this.

Glyph	*Name*	*Comments*	*Examples*
	ARABIC SIGN SALLALLAHOU ALAYHE WASALLAM	Represents *sallallahu alayhe wasallam* 'may God's peace and blessings be upon him'	See figures 2, 6, 11, 12, 14, 18, 31
	ARABIC SIGN ALAYHE ASSALAM	Represents *alayhe assalam* 'upon him be peace'	See figures 3, 11, 17, 21, 22, 23
	ARABIC SIGN RAHMATULLAH ALAYHE	Represents *rahmatullah alayhe* 'may God have mercy upon him'	See figures 3, 7, 23, 29
	ARABIC SIGN RADI ALLAHU ANHU	Represents *radi allahu 'anhu* 'may God be pleased with him'	See figures 3, 6, 12, 15, 18, 19, 21, 28
	ARABIC SIGN NOM DE PLUME	Sign placed over the name or nom-de-plume of a poet, or in some writings used to mark all proper names	See figures 4, 14, 15, 16, 18, 25, 29, 30

5. Date and number signs

There are three special signs written in association with numbers in Arabic script that are not currently present in the UCS repertoire.

The first is a separator used between the (numeric) date and the month name when writing out a date. Figure 1 shows that this sign is distinct from U+002F SOLIDUS (used, for example, as a separator in currency amounts).

The second signals the beginning of a number, and is written below the digits of the number, and the third indicates a year (i.e., as part of a date). This sign also combines below the digits of the number. Its appearance is a vestigial form of the Arabic word for "year" /sanatu/ (SEEN NOON TEH-MARBUTA), but is now a sign in its

own right, widely used to mark a numeric year even in non-Arabic languages where the Arabic word would not be known.

Both the year and number signs need to be able to combine with multi-digit numbers, not just with a single base character. This is also true for the character U+06DD ARABIC END OF AYAH already present in the UCS repertoire. It is not currently clear to implementers exactly how such a character should behave. The proposed new character COMBINING GRAPHEME JOINER may be relevant here, but it seems most unfortunate to require users to insert grapheme joiners between the digits of a multi-digit year or Ayah number.

All these signs should be encoded in the UCS, as they are in widespread use.

Some additional statement clarifying how these marks, including U+06DD as well as the newly proposed characters ARABIC NUMBER SIGN and ARABIC YEAR SIGN, combine with digit *sequences* would be helpful.

Glyph	*Name*	*Comments*	*Examples*
	ARABIC DATE SEPARATOR	This is a specific character used in writing dates; distinct from SOLIDUS	See figures 1, 18
	ARABIC NUMBER SIGN	Most logically, the sign should precede the digits in the text stream, but Unicode principles require it (as a combining mark) to follow them	See figures 2, 14, 22, 29
	ARABIC YEAR SIGN	Same issue as ARABIC NUMBER SIGN	See figures 1, 5, 7, 18, 31

6. Poetic verse sign and footnote marker

There is a special symbol often used to mark the beginning of a poetic verse. As this symbol is peculiar to Arabic script, it is more appropriately encoded in the Arabic block (like the specifically Arabic punctuation marks) rather than among general symbols or dingbats.

A similar sign is also used in Arabic script as a footnote marker, in conjunction with the footnote number (it combines with the digits similarly to ARABIC NUMBER SIGN). Although the poetic verse and footnote signs look similar, the one is a symbol and the other a combining character. For this reason, it is not possible to unify the characters.

The footnote marker may combine with multiple base characters, just like the number and year signs discussed above.

Glyph	*Name*	*Comments*	*Examples*
	ARABIC POETIC VERSE SIGN	Distinct from ARABIC FOOTNOTE MARKER (below) because this is a spacing symbol	See figures 8, 18, 20
	ARABIC FOOTNOTE MARKER	Distinct from ARABIC POETIC VERSE SIGN (above) because this is a combining character	See figures 2, 10, 30

7. Summary of requested characters

The proposed additions to the UCS are summarized here, together with the important Unicode properties needed for each character. (The remaining fields of the Unicode Character Database will be empty for all these characters.)

Canonical combining classes are not necessarily easy to choose; I believe that the classes assigned to a number of the existing Arabic-block marks are unfortunate, in that they lead to a canonical ordering that does not make good linguistic sense. As modifying them would raise serious compatibility issues, however, we are not proposing any change to the properties of existing characters.

In the proposed classes below, I 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 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 annotations should 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.
- U+06DD: mark combines with maximal preceding sequence of digits, not just a single character.

Representative glyph	*Suggested USV*	*Character name*	*General category*	*Combining class*	*Bidi category*
	U+0656	ARABIC SUBSCRIPT ALEF	Mn	37	NSM
	U+0657	ARABIC TURNED DAMMA	Mn	38	NSM
	U+0658	ARABIC NASALIZATION MARK	Mn	7	NSM
	U+0659	ARABIC SIGN SALLALLAHOU ALAYHE WASALLAM	Mn	250	NSM
	U+065A	ARABIC SIGN ALAYHE ASSALAM	Mn	250	NSM
	U+065B	ARABIC SIGN RAHMATULLAH ALAYHE	Mn	250	NSM
	U+065C	ARABIC SIGN RADI ALLAHU ANHU	Mn	250	NSM
	U+065D	ARABIC SIGN NOM DE PLUME	Mn	250	NSM
	U+065E	ARABIC NUMBER SIGN	Me	0	NSM
	U+065F	ARABIC YEAR SIGN	Me	0	NSM

؍	U+060D	ARABIC DATE SEPARATOR	Po	0	AL
۞	U+060E	ARABIC POETIC VERSE SIGN	So	0	ON
۟	U+060F	ARABIC FOOTNOTE MARKER	Me	0	NSM

II. Examples of usage

See Appendix (provided as a separate document) for Figures 1–36. Note that these examples show that all the proposed characters predate the DTP era, and represent established practice. Proposal L2/00-135 shows that current font/DTP vendors also wish to support them in data processing and interchange.

III. References

A. Sources of character examples

[1] Mohammad Abd-al-Rahman Barker and Aqil Khan Mengal, 1969. A Course in Baluchi, Vol. 2. Institute of Islamic Studies, McGill University, Montreal.

[2] Ali Asghar Chaudhry, 1992. ہمارے آقا حضُور (Our Noble Master). Maktabah Tamir-e-Insaniat, Lahore.

[3] Rahman Firaz, 1983. نیا دِن (Naya Din [New Day], Urdu adult literacy course). Nirali Kitaben, Lahore.

[4] Sayad Hashmi, 2000. سَیَد گنج (Sayad Ganj [Sayad's Treasury], Balochi Dictionary). Sayad Hashmi Academy, Karachi.

[5] Maulana Abdul Aziz Hazarvi, 1987. بڑا قصص ُالانبیاء (Big Stories of the Prophets). Maktabah Aziziah, Karachi.

[6] Qamar Ahmad Ali Khan, n.d. آؤ بچّو اُردُو سیکھیں (Come Child, Learn Urdu), Book 1. Ferozsons Pvt. Ltd., Karachi.

[7] ———, n.d. آؤ بچّو اُردُو سیکھیں (Come Child, Learn Urdu), Book 2. Ferozsons Pvt. Ltd., Karachi.

[8] Mir Mitha Khan Marri and Ghaus Bakhsh Sabir, eds., 1985. مَن پاکِستانی آں (I am Pakistani). Pakistan Children's Academy, Balochistan.

[9] Maulvi Khair Mohammad Nadwi, December 1987. ماہتاك سوغات (Saughat [Gift] Magazine). Maktabah Saughat, Karachi.

[10] ———, July 1989. ماہتاك سوغات (Saughat [Gift] Magazine). Maktabah Saughat, Karachi.

[11] Pakistan Bible Society, 1992. کِتابِ مُقدّس (The Holy Bible in Urdu, Revised Version with References). Pakistan Bible Society, Lahore.

[12] Ghani Parwaz, 1997. لبزانکی شرگِدارِی (Literary Criticism). Balochi Academy, Quetta.

[13] Alama Qazi Abdus Samad Sarbazi and Maulana Khair Mohammad Nadwi, 1986. قُرآن مَجید ترجمه و تفسیر بلوچی (The Great Qur'an, Translation and Explanation in Balochi). Maktabah Ishaqiah, Karachi.

[14] Muhammad Iqbal Siddiqi, 1987. Ninety Nine Names of Allah. Kazi Publications, Lahore.

[15] Abdul Ghani Warisi. Abdur Rashid Warisi and Aslam Ahmad Hadi Qureshi, n.d. پیاری نعتیں (Lovely Eulogies). Jehangir Book Depot, Lahore.

[16] القرآن الکریم (The Noble Qur'an), 1405 A.H. (1984-5). King Fahd Holy Qur'an Printing Complex, Medina.

B. Related discussion documents

[17] Kamal Mansour, Thomas Milo and Roozbeh Pournader, August 2001. Email discussion of L2/01-304 on the Unicore list.

[18] Thomas Milo, 2001. Variants in Arabic script – related to L2/01-304 (L2/01-325).

[19] Thomas Milo, 2001. Some comments on the Arabic block in Unicode (L2/01-329).

[20] Paul Nelson, Ashhar Farhan, Arif Hisam, Kashif Hisam and John Clews, 2000. Proposal to Add Urdu Epethit *[sic]* and Abbreviation Diacritics to Arabic Block. L2/00-135.