L2/03-176

Proposal to encode Jawi and Moroccan Arabic GAF characters

Date: June 3, 2003
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 encode Jawi and Moroccan Arabic GAF characters
2.	Requester's name	SIL International (contacts: Jonathan Kew, Peter Constable)
3.	Requester type	Expert contribution
4.	Submission date	June 3, 2003
5.	Requester's reference	
6a.	Completion	This is a complete proposal.
6b.	More information to be provided?	Only as required for clarification.

B. Technical — General

1a.	New script? Name?	No
1b.	Addition of characters to existing block? Name?	Yes — Arabic
2.	Number of characters in proposal	2
3.	Proposed category	A
4.	Proposed level of implementation and rationale	1 (no 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?	No
2a.	Has contact been made to members of the user community?	No
2b.	With whom?	N/A
3.	Information on the user community for the proposed characters is included?	Yes
4.	The context of use for the proposed characters	Publications in Jawi (Malay), north African Arabic, Amazigh languages.
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 current use.
7.	Should the proposed characters be kept together in a contiguous range?	Together with existing Arabic characters.
8a.	Can any of the proposed characters be considered a presentation form of an existing character or character sequence?	No (but see L2/03-154, and discussion below).
8b.	Rationale for inclusion?	N/A
9a.	Can any of the proposed characters be considered to be similar (in appearance or function) to an existing character?	See discussion below.
9b.	Rationale for inclusion?	See below.
10.	Does the proposal include the use of combining characters and/or use of composite sequences?	No
11.	Does the proposal contain characters with any special properties?	Characters have right-to-left (AL) directionality.

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

The following two characters are proposed as additions to the UCS repertoire. They could possibly be viewed as variants of U+06AC and U+06AD respectively, but the different Arabic joining groups involved, as well as the fact that shapes based on ك and ک are not interchangeable for all purposes or in all languages, suggests that such unification is not the most appropriate course.

Both characters have similar properties: general category Lo, bidirectional type AL, Arabic shaping group GAF, combining class 0, not mirrored, no numeric value.

Glyph	*Code*	*Character name*	*Notes*
ڪ	065D	ARABIC LETTER KEHEH WITH ONE DOT ABOVE	Preferred over 06AC for /g/ in Jawi
ڮ	065E	ARABIC LETTER KEHEH WITH THREE DOTS ABOVE	North African /g/

Note that if the Arabic-script modifier marks proposed in L2/03-154 (Kew et al 2003) are encoded, it will then be possible to represent these letters using sequences of *base + modifier,* with the base character being U+06A9 ARABIC LETTER KEHEH, and there will be no need to encode them as individual characters.

II. Rationale

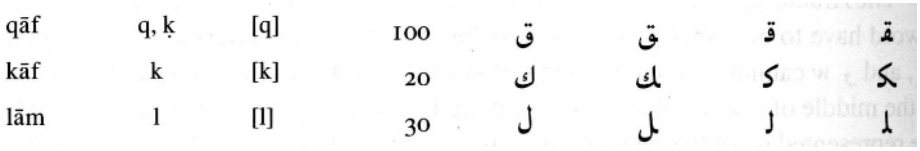
1. Background

The Arabic letter *kaf,* nominally representing a /k/ phoneme, has several clearly distinct graphical forms. These derive from varying calligraphic traditions, and Arabic speakers understand them to be mere variations in the style of writing a single letter. As such, all can be represented by a single encoded character, U+0643 ARABIC LETTER KAF. Representative glyphs showing the three major forms of this letter, labeled A, B, and C, are shown in figure 1:

A B C

Figure 1: Three forms of the Arabic letter *kaf*

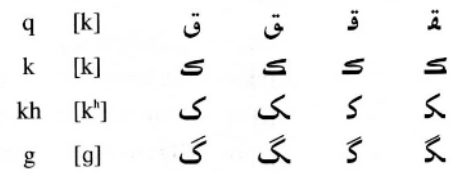
Form A (ك) is the form most commonly used in current text fonts, and is appropriately chosen for the representative glyph in the Unicode standard; it is also the form seen in typical charts of the Arabic alphabet, such as figure 2:



qāf	q, ḳ	[q]	100	ق	ق	ـق	قـ
kāf	k	[k]	20	ك	ك	ـك	كـ
lām	l	[l]	30	ل	ل	ـل	لـ

Figure 2: From chart of the Arabic alphabet, Daniels & Bright (1996), page 560.

However, where the Arabic script has been adopted for writing non-Arabic languages, variations in form that in Arabic were free variation or stylistic variants have sometimes been co-opted to make meaningful distinctions that merit encoding as separate characters. A clear example of this can be seen in Sindhi, where two forms of *kaf* are used as separate letters of the alphabet. The important phonemic distinction between /k/ (unaspirated) and /k^h/ (aspirated) is represented by using form C (ڪ) for /k/ and form B (ک) for /k^h/, as shown in figure 3; the typical Arabic form A (ك) is not used in Sindhi.



q	[k]	ق	ق	ـق	قـ
k	[k]	ڪ	ڪ	ـڪ	ڪـ
kh	[k^h]	ک	ک	ـک	کـ
g	[g]	گ	گ	ـگ	گـ

Figure 3: From chart of the Sindhi alphabet, Daniels & Bright (1996), page 757.

To support the Sindhi usage (and other similar situations), we note that Unicode encodes the three *kaf* forms separately as distinct characters:

Code	*Glyph*	*Name*	*Joining*
0643	ك	ARABIC LETTER KAF	KAF
06A9	ک	ARABIC LETTER KEHEH	GAF
06AA	ڪ	ARABIC LETTER SWASH KAF	SWASH KAF

Figure 4: Forms of Arabic *kaf* encoded in Unicode 4.0

(The name KEHEH used for U+06A9 is probably an attempt to transcribe the Sindhi name for the letter representing aspirated /k^h/.) Figure 4 also shows that Unicode assigns these three characters to distinct joining groups, reflecting the fact that they are substantially different graphical forms and must each be shaped according to a different pattern.

A similar example of the disunification of Arabic glyph variants to become distinct characters when used for another language can be seen in the Urdu usage of the letter *heh.* Here, a contrast between ہ and ھ (and their related linking forms) is consistently used to distinguish the independent letter /h/, written with ہ, from aspiration of plosives and affricates, written with ھ. Arabic speakers would consider these variants of a single letter, but a clear distinction must be encoded for some other languages (and is therefore supported in Unicode).

Yet another example occurs with *yeh*: to an Arabic speaker, the form ے is merely a calligraphic variant of the letter ي (or ى). But in Urdu, the form ے has been adopted to write the vowel /e/, while the form ى represents /i/. This distinction must be encoded, and so Unicode includes U+06D2 ے as a separate character.

So we see that where there are clearly distinct graphical forms in existence for an Arabic letter, it may well be appropriate to encode these forms separately. The fact that they originate as different calligraphic styles of a single letter in the Arabic language does not mean that this interpretation is adequate for all languages and regions.

Given that Unicode encodes these three forms of *kaf* separately, it seems appropriate to treat modified forms of *kaf* in a similar way. Where additional letters have been created by adding dots or other marks to an underlying *kaf*, this has often been done to one specific form of the letter (or in Unicode terms, to one of the three characters U+0643, 06A9, 06AA), and substitution of a different base form may not be at all acceptable. This is clear, for example, in the case of the Persian and Urdu /g/, written as U+06AF گ; the added 'bar' that creates the letter *gaf* can only be added to form B of the *kaf.*

2. Jawi GAF

In the Jawi script (Arabic script used to write Malay), the /g/ sound is written using a *kaf* with one dot above. Such a character is encoded in Unicode at U+06AC (ڬ), with representative glyph based on form A of *kaf.* However, the Jawi *gaf* is properly based on form B, not form A:

ḳ	[k]	ق	ق	ﻗ	ﻘ
k	[k]	ك	ك	ﻛ	ﻜ
g	[g]	ک	ک	ﮔ	ﮕ
l	[l]	ل	ل	ﻟ	ﻠ

Figure 5: From chart of the Jawi alphabet, Daniels & Bright (1996), page 761.

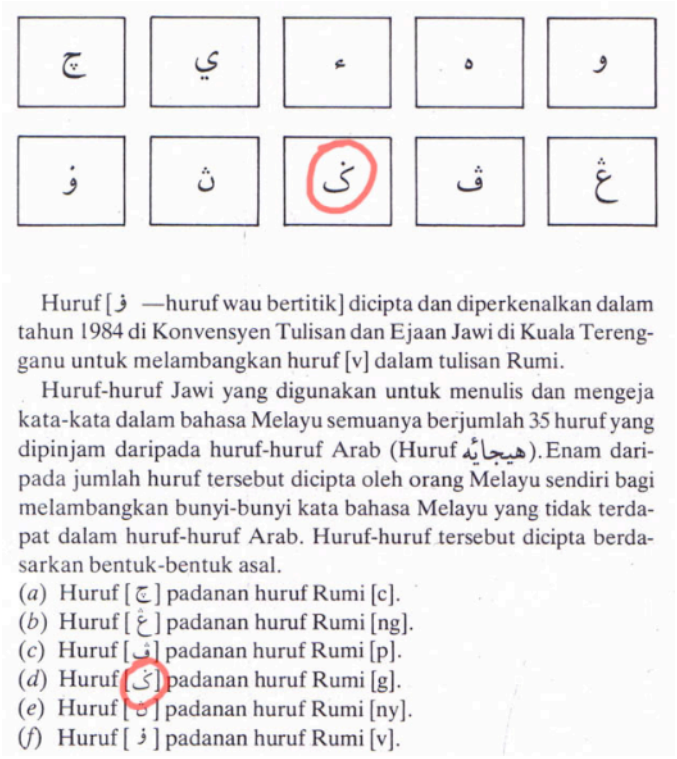
Note in figure 5 that form B is used as the basis for the /g/ character, despite the fact that form A is used for /k/. This is a characteristic of Jawi writing, and not an artifact of this particular book's typography. Comparing the chart for Uighur, found on the previous page, we see that in some cases form A is used as the base for a modified *kaf*; the use of form B in the Jawi chart is no accident.

q	[q]	ق	ق	ﻗ	ﻘ
k	[k]	ك	ك	ﻛ	ﻜ
g	[g]	گ	گ	ﮔ	ﮕ
ng	[ŋ, n]	ڭ	ڭ	ﯓ	ﯔ
l	[l]	ل	ل	ﻟ	ﻠ

Figure 6: From chart of the Uighur alphabet, Daniels & Bright (1996), page 760.

Figure 6 shows a form A *kaf* with three dots above, used for the /ŋ/ sound. Comparing this with the Jawi chart, we see that new letters based on *kaf* may involve a deliberate choice of one of the three forms of *kaf*, which may not be interchangeable or treated as glyph variants in this context.

All Jawi sources I have seen show the use of form B as the basis of the /g/, even though form A is commonly used for /k/. Figure 7 is taken from an introduction to the Jawi script published in Malaysia.


چ ي ء ه و

ڤ ڽ ڬ ف غ

Huruf [ڤ —huruf wau bertitik] dicipta dan diperkenalkan dalam tahun 1984 di Konvensyen Tulisan dan Ejaan Jawi di Kuala Terengganu untuk melambangkan huruf [v] dalam tulisan Rumi.

Huruf-huruf Jawi yang digunakan untuk menulis dan mengeja kata-kata dalam bahasa Melayu semuanya berjumlah 35 huruf yang dipinjam daripada huruf-huruf Arab (Huruf هيجائيه). Enam daripada jumlah huruf tersebut dicipta oleh orang Melayu sendiri bagi melambangkan bunyi-bunyi kata bahasa Melayu yang tidak terdapat dalam huruf-huruf Arab. Huruf-huruf tersebut dicipta berdasarkan bentuk-bentuk asal.

(*a*) Huruf [چ] padanan huruf Rumi [c].
(*b*) Huruf [غ] padanan huruf Rumi [ng].
(*c*) Huruf [ف] padanan huruf Rumi [p].
(*d*) Huruf [ڬ] padanan huruf Rumi [g].
(*e*) Huruf [ڽ] padanan huruf Rumi [ny].
(*f*) Huruf [ڤ] padanan huruf Rumi [v].

Figure 7: From Muhani (1998), page 6.

It is clear from the notes in the names list that the Unicode character U+06AC (ﻙ) was encoded with the intent that it be used for Jawi /g/; however, we see that the representative glyph shown in the code charts and the joining group listed in *ArabicShaping.txt* are inappropriate for this purpose.

One possible response would be to change the glyph and the joining group of U+06AC to those expected in Jawi, thus making this character suitable for its originally-intended purpose. However, given the tendency, especially once modifying marks are added, for users to make a clear distinction between the different forms of *kaf*, not considering them merely as glyph variants of a single character, this comes dangerously close to changing the fundamental identity of the character. Moreover, there can be no assurance that a character having the specific form A with a dot above has *not* been deliberately used in some context, given that it exists in the current standard.

The fact that the Arabic joining group is considered a normative property of the Unicode character also weighs against the position that ﻙ and ڬ could be considered variants of the same character; they must necessarily have different joining groups. It is therefore proposed that a new character ARABIC LETTER KEHEH WITH ONE DOT ABOVE should be encoded, and a note added to U+06AC indicating that the new character is preferred for old Malay.

3. Moroccan GAF

Although standard Arabic does not write a /g/ sound, in Morocco the use of a form B *kaf* with three dots above is well established as the letter representing /g/. Published literature is generally in standard Arabic, and as such does not use this letter, but it is seen in other situations such as road signs, product labels, etc. It is also used in writing the Amazigh languages of Morocco. The following photographs show examples of this letter used in Moroccan Arabic:

Figure 8: Street name in Arabic and Latin scripts.

Figure 9: Petrol pump, labeled ‘gasoil’ in Arabic script.

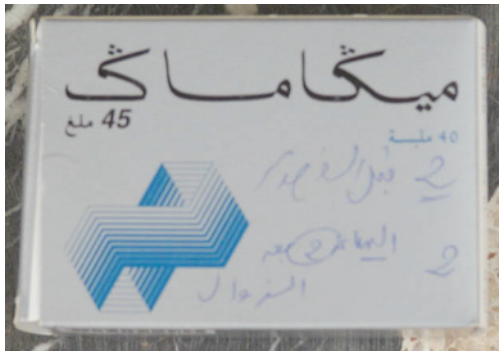


Figure 10: Medicine package ‘Megamag’, showing that the *gaf* is based on *kaf* form B.

Figure 11 is taken from the Royal Moroccan Academy’s new Amazigh dictionary, and shows the letter in initial, medial, and final forms.

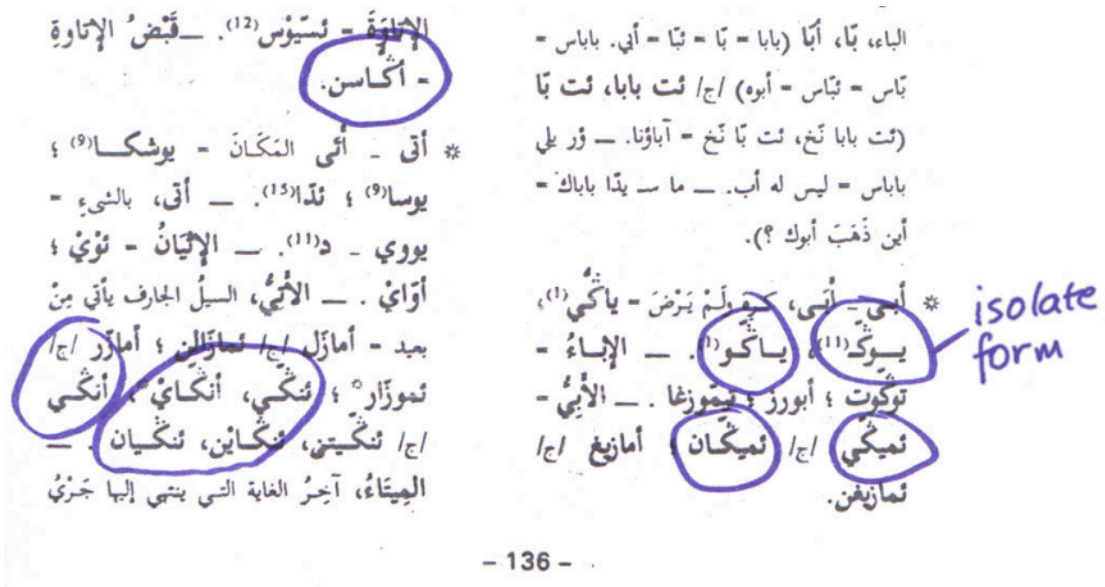


Figure 11: From Shafiq (1996), page 136; note the “makeshift” isolated *gaf*.

Close examination of figure 11 shows that the author did not have an appropriate character available, and therefore used a standard *kaf* and added the three dots by hand. The isolated form is of particular interest, as here the author deliberately added a *kashida* (extender) character in order to achieve an approximation of *kaf* form B, in preference to simply adding the dots to the form A *kaf* which would otherwise have appeared.

A similar form is seen in Mauritanian texts, where a *kaf* written in shape B with three dots added is also used to represent /g/:

ḍhar ظهر	The 'black' way (*jaamba*) الطريق الكحلاء	*gnaydiiya* الكْنيديّه	The 'white' way (*jaamba*) الطريق البيضاء
kar كر	[1]*entemaas* انتماس	*enweffal* انوفّل	[1]*mekka mūsa* مكّة موس
	[2]*sayni kar (ma yuḳarraṣ)* سيني كر (ما يخرّص)		[2]*el-faayez* الفايز
faaġu فاغ	[1]*tenačču uga* تنچّوگه	*šš baar* الشّبار	[1]*ssruuzi ('arraay essruuza)* السّروزي (عرّاي السّروزه)
	[2]*sayni faaġu* سيني فاغ	or *faaġu lekbiir* فاغ الكبير	[2]*et-teḥraar (el-ḥurr)* التّحرار (الحرّ)
siññiima سنّيمه	[1]*siññiimat ḥayba*[3] سنّيمة هيب	('black') *el-mawṣṭi* الموسطي	[1]*et-teḥzaam*[4] التّحزام
	[2]*meqaččuuga* (Ḥawd) مقچّوگه		[2]*eññaama*[5] انّيامه
	or *lebyaaḍ* لبياظ	('white') *menčalla* منچْلّه	[2]*rrbaabi*[6] الرّبابي
	or *etbaybi* (Trārza) اتبيبي		(a) *leggetri*[6] لگّتري
			(b) *čaynna* چينّه
			(c) *liyyin* ليّن
baygi or *lebtayt* بيڭي	*baygi* بيڭي	*el-muḳaalef* المخالف	*el-'itiig* العتيڭ
	intensive form *a'ḍḍaal* اعضّال	or *baygi jjraad* بيڭي الجّراد	
		or *baygi lekbiir* بيڭي الكبير	

[1] Entry (*dḳuul* دخول) brought about through tightening the strings of the *tidīnīt*.
[2] Subsidiary (*rrdiif* الرّديف) brought about through loosening the strings of the *tidīnīt*.
[3] Also called in western Mauritania *zzraag* الزّراگ.
[4] Included in *faaġu* by some *iggāwen*.
[5] Not subsidiary to those who include *et-teḥzaam* in *faaġu*.
[6] A blending of white and black.

Figure 12: From Norris (1968), page 73. Note contrasting basic shape used for final *kaf* and *gaf*.

Here again, the printer has been forced to make do with a limited selection of available glyphs, and the results are instructive. The final and isolated forms of *gaf* are constructed by adding the 'tail' of a *kaf* (intended for use in building form A) to a medial or initial *kaf* with three dots added, giving a result that resembles form B in having the added bar on top, but also has the 'flourish' typical of form A. The importance of the 'form B-ness' of this letter is evident when we note the trouble that has been taken to build it, in contrast to the simple form A used for *kaf* itself (highlighted in blue in figure 12).

In initial and medial joined forms, this letter would be visually identical to U+06AD ڭ ARABIC LETTER NG. However, in final and isolated forms the difference is clear; Moroccan /g/ is consistently based on form B of *kaf*, even though form A is commonly used for the letter *kaf* itself. Given the clear distinction between these forms in user's minds; the fact that the forms ڭ and ڮ are not considered interchangeable, even where ك and ک are understood to be variants of the same letter *kaf*; and the different joining groups required, it is proposed that a new character ARABIC LETTER KEHEH WITH THREE DOTS ABOVE should be encoded in the UCS to represent the Moroccan/Amazigh *gaf*.

III. References

Daniels, Peter T. and William Bright (eds). 1996. *The world's writing systems*. New York/Oxford: OUP.

Kew, Jonathan, Mark Davis and Kamal Mansour. 2003. *Proposal to encode productive Arabic-script modifier marks*. L2/03-154.

Muhani, Hj. Abdul Ghani. 1988. *Teman pelajar Jawi*. Petaling Jaya (Malaysia): Fajar Bakti.

Norris, H. T. 1968. *Shinqiti folk literature and song*. Oxford: Clarendon Press.

Shafiq, Muḥammad (ed). 1990. *al-Muʿjam al-ʿArabi al-Amazighi*. Rabat: Akadimiyat al-Mamlakah al-Maghribiyah [Royal Moroccan Academy].