ISO/IEC JTC 1/SC 2/WG 2
PROPOSAL SUMMARY FORM TO ACCOMPANY SUBMISSIONS
FOR ADDITIONS TO THE REPERTOIRE OF ISO/IEC 10646[1]
Please fill all the sections A, B and C below.
(Please read Principles and Procedures Document for guidelines and details before filling this form.)
See http://www.dkuug.dk/JTC1/WG2/docs/summaryform.html **for latest *Form*.**
See http://www.dkuug.dk/JTC1/WG2/docs/principles.html **for latest *Principles and Procedures* document.**
See http://www.dkuug.dk/JTC1/WG2/docs/roadmaps.html **for latest roadmaps.**

A. Administrative

1. **Title:** Proposal to encode additional Arabic-script characters
2. Requester's name: INCITS/L2; Unicode Technical Committee; Jonathan Kew, SIL International
3. Requester type (Member body/Liaison/Individual contribution): Member; Liaison; expert contribution
4. Submission date: 2003-07-10
5. Requester's reference (if applicable): L2/02-274; L2/03-159; L2/03-168; L2/03-176; L2/03-210
6. (Choose one of the following:)
   This is a complete proposal: Yes
   or, More information will be provided later:

B. Technical - General

1. (Choose one of the following:)
   a. This proposal is for a new script (set of characters): No
      Proposed name of script:
   b. The proposal is for addition of character(s) to an existing block: Yes
      Name of the existing block: Arabic, and proposed Arabic Supplementary block
2. Number of characters in proposal: 30
3. Proposed category (see section II, Character Categories): A
4. Proposed Level of Implementation (1, 2 or 3) (see clause 14, ISO/IEC 10646-1: 2000): 2
   Is a rationale provided for the choice? Yes
   If Yes, reference: Includes combining marks
5. Is a repertoire including character names provided? Yes
   a. If YES, are the names in accordance with the 'character naming guidelines in Annex L of ISO/IEC 10646-1: 2000? Yes
   b. Are the character shapes attached in a legible form suitable for review? Yes
6. Who will provide the appropriate computerized font (ordered preference: True Type, or PostScript format) for publishing the standard? Jonathan Kew, SIL International
   If available now, identify source(s) for the font (include address, e-mail, ftp-site, etc.) and indicate the tools used:
   jonathan_kew@sil.org
   TrueType font generated with FontLab 4.5
7. References:
   a. Are references (to other character sets, dictionaries, descriptive texts etc.) provided? Yes
   b. Are published examples of use (such as samples from newspapers, magazines, or other sources) of proposed characters attached? Yes
8. Special encoding issues:
   Does the proposal address other aspects of character data processing (if applicable) such as input, presentation, sorting, searching, indexing, transliteration etc. (if yes please enclose information)?
   Yes: suggested Unicode character properties are included
9. Additional Information:
Submitters are invited to provide any additional information about Properties of the proposed Character(s) or Script that will assist in correct understanding of and correct linguistic processing of the proposed character(s) or script. Examples of such properties are: Casing information, Numeric information, Currency information, Display behaviour information such as line breaks, widths etc., Combining behaviour, Spacing behaviour, Directional behaviour, Default Collation behaviour, relevance in Mark Up contexts, Compatibility equivalence and other Unicode normalization related information. See the Unicode standard at http://www.unicode.org for such information on other scripts. Also see http://www.unicode.org/Public/UNIDATA/UnicodeCharacterDatabase.html and associated Unicode Technical Reports for information needed for consideration by the Unicode Technical Committee for inclusion in the Unicode Standard.

C. Technical - Justification

1. Has this proposal for addition of character(s) been submitted before? Not to WG2
   If YES explain (but characters approved for encoding at UTC #95, June 2003)
2. Has contact been made to members of the user community (for example: National Body, user groups of the script or characters, other experts, etc.)? Yes
   If YES, with whom? Local communities, linguists, NGOs working in S. Asia & N. Africa
   If YES, available relevant documents: See §4 below
3. Information on the user community for the proposed characters (for example: size, demographics, information technology use, or publishing use) is included? Yes
   Reference: See §3 below
4. The context of use for the proposed characters (type of use; common or rare) Various
   Reference: See §3 below
5. Are the proposed characters in current use by the user community? Yes
   If YES, where? Reference: South Asia, North African countries
6. After giving due considerations to the principles in *Principles and Procedures document* (a WG 2 standing document) must the proposed characters be entirely in the BMP? Yes
   If YES, is a rationale provided? Yes
   If YES, reference: Extensions to BMP Arabic repertoire, see §3 below
7. Should the proposed characters be kept together in a contiguous range (rather than being scattered)? Yes
8. Can any of the proposed characters be considered a presentation form of an existing character or character sequence? No
   If YES, is a rationale for its inclusion provided?
   If YES, reference:
9. Can any of the proposed characters be encoded using a composed character sequence of either existing characters or other proposed characters? No
   If YES, is a rationale for its inclusion provided?
   If YES, reference:
10. Can any of the proposed character(s) be considered to be similar (in appearance or function) to an existing character? Yes
    If YES, is a rationale for its inclusion provided? Yes
    If YES, reference: See §3.2 below
11. Does the proposal include use of combining characters and/or use of composite sequences (see clauses 4.12 and 4.14 in ISO/IEC 10646-1: 2000)? Yes
    If YES, is a rationale for such use provided? Yes
    If YES, reference: Arabic-script vowels (see §3.1.2 below) are combining marks
    Is a list of composite sequences and their corresponding glyph images (graphic symbols) provided? No
    If YES, reference:
12. Does the proposal contain characters with any special properties such as control function or similar semantics? No
    If YES, describe in detail (include attachment if necessary)
13. Does the proposal contain any Ideographic compatibility character(s)? No
    If YES, is the equivalent corresponding unified ideographic character(s) identified?
    If YES, reference:

[1]Form number: N2352-F (Original 1994-10-14; Revised 1995-01, 1995-04, 1996-04, 1996-08, 1999-03, 2001-05, 2001-09)

Submitter's Responsibilities

The national body or liaison organization (or any other organization or an individual) proposing new character(s) or a new script shall provide:

1. Proposed category for the script or character(s), character name(s), and description of usage.
2. Justification for the category and name(s).
3. A representative glyph(s) image on paper:
   If the proposed glyph image is similar to a glyph image of a previously encoded ISO/IEC 10646 character, then additional justification for encoding the new character shall be provided.
   Note: Any proposal that suggests that one or more of such variant forms is actually a distinct character requiring separate encoding, should provide detailed, printed evidence that there is actual, contrastive use of the variant form(s). It is insufficient for a proposal to claim a requirement to encode as characters in the Standard, glyphic forms which happen to occur in another character encoding that did not follow the Character-Glyph Model that guides the choice of appropriate characters for encoding in ISO/IEC 10646.
   Note: WG 2 has resolved in Resolution M38.12 not to add any more Arabic presentation forms to the standard and suggests users to employ appropriate input methods, rendering and font technologies to meet the user requirements.
4. Mappings to accepted sources, for example, other standards, dictionaries, accessible published materials.
5. Computerized/camera-ready font:
   Prior to the preparation of the final text of the next amendment or version of the standard a suitable computerized font (camera-ready font) will be needed. Camera-ready copy is mandatory for final text of any pDAMs before the next revision. Ordered preference of the fonts is True Type or PostScript format. The minimum design resolution for the fontis 96 by 96 dots matrix, for presentation at or near 22 points in print size.
6. List of all the parties consulted.
7. Equivalent glyph images:
   If the submission intends using composite sequences of proposed or existing combining and non-combining characters, a list consisting of each composite sequence and its corresponding glyph image shall be provided to better understand the intended use.
8. Compatibility equivalents:
   If the submission includes compatibility ideographic characters, identify the equivalent unified CJK Ideograph character(s).
9. Any additional information that will assist in correct understanding of the different characteristics and linguistic processing of the proposed character(s) or script.

1. Proposed character additions (shaded cells)

2. Names list for character additions

Listing of file *ArabicUTC95-names.txt*

```
; additions to the 0600 Arabic block
;
@@   0600   Arabic   06FF
;
@       Punctuation
061E   ARABIC TRIPLE DOT PUNCTUATION MARK
;
@       Other combining marks
;
;See Everson/Pournader's proposal L2/03-133R or N2581R2
;0659   ARABIC ZWARAKAY
;    * Pashto
;
065A   ARABIC VOWEL SIGN SMALL V ABOVE
    * African languages
065B   ARABIC VOWEL SIGN INVERTED SMALL V ABOVE
    * African languages
065C   ARABIC VOWEL SIGN DOT BELOW
    * African languages
;
; new Arabic Supplementary block
;
@@   0750   Arabic Supplementary   077F
;
@       Extended Arabic letters
@+       These are primarily used in Arabic-script orthographies of African languages.
0750   ARABIC LETTER BEH WITH THREE DOTS HORIZONTALLY BELOW
0751   ARABIC LETTER BEH WITH DOT BELOW AND THREE DOTS ABOVE
0752   ARABIC LETTER BEH WITH THREE DOTS POINTING UPWARDS BELOW
0753   ARABIC LETTER BEH WITH THREE DOTS POINTING UPWARDS BELOW AND TWO DOTS ABOVE
0754   ARABIC LETTER BEH WITH TWO DOTS BELOW AND DOT ABOVE
0755   ARABIC LETTER BEH WITH INVERTED SMALL V BELOW
0756   ARABIC LETTER BEH WITH SMALL V
0757   ARABIC LETTER HAH WITH TWO DOTS ABOVE
0758   ARABIC LETTER HAH WITH THREE DOTS POINTING UPWARDS BELOW
0759   ARABIC LETTER DAL WITH TWO DOTS VERTICALLY BELOW AND SMALL TAH
    * Saraiki
075A   ARABIC LETTER DAL WITH INVERTED SMALL V BELOW
075B   ARABIC LETTER REH WITH STROKE THROUGH
075C   ARABIC LETTER SEEN WITH FOUR DOTS ABOVE
    * Shina
075D   ARABIC LETTER AIN WITH TWO DOTS ABOVE
075E   ARABIC LETTER AIN WITH THREE DOTS POINTING DOWNWARDS ABOVE
075F   ARABIC LETTER AIN WITH TWO DOTS VERTICALLY ABOVE
0760   ARABIC LETTER FEH WITH TWO DOTS BELOW
0761   ARABIC LETTER FEH WITH THREE DOTS POINTING UPWARDS BELOW
0762   ARABIC LETTER KEHEH WITH DOT ABOVE
    * old Malay, preferred to 06AC
    x (kaf with dot above - 06AC)
0763   ARABIC LETTER KEHEH WITH THREE DOTS ABOVE
    * Moroccan Arabic, Amazigh
    x (arabic letter ng - 06AD)
0764   ARABIC LETTER KEHEH WITH THREE DOTS POINTING UPWARDS BELOW
0765   ARABIC LETTER MEEM WITH DOT ABOVE
0766   ARABIC LETTER MEEM WITH DOT BELOW
    * Maba
0767   ARABIC LETTER NOON WITH TWO DOTS BELOW AND DOT ABOVE
0768   ARABIC LETTER NOON WITH DOT ABOVE AND SMALL TAH
    * Saraiki, Pathwari
0769   ARABIC LETTER NOON WITH SMALL V
    * Gojri
```

3. Discussion of the proposed additions

The proposed characters can be considered in several categories: a collection of letters that have been used to extend Arabic script for various African languages; two *gaf* letters that merit individual discussion; several additional letters used in South Asian languages; and a punctuation character traditionally used when writing African languages in Arabic script.

3.1 North African extensions to Arabic script

The principal source of information concerning many of the characters proposed for encoding here is Chtatou (1992). By way of general information on the languages involved, Chtatou writes:

> *Fulfulde*
>
> This language is also known by other names, mainly: Fula, Peul, Pular and Pulaar, and is spoken in many countries ... Fulfulde covers a larger geographical area than any other African language and, as a result, is considered as one of the principal languages of Africa: it is spoken by between 12 and 15 million people.
>
> In the field of education, Fulfulde is used as a language of instruction in Guinea and Nigeria and there are pilot-projects considering its use in the Gambia, Mauritania and Niger. In addition, it is used as a language of instruction at secondary school level for the first two years in Guinea and at the university level in Nigeria. The language in question is used by the press and, as a result, two monthlies appear in it: one in Mali with a circulation of 500 copies and another in Niger with 3000 copies.
>
> *Hausa*
>
> Geographically speaking, Hausa is less spread out than Fulfulde; it is one of the principal languages of Africa, spoken by 40 million people ... An intense literary activity involving this language has been signaled mainly in the field of fiction and poetry. At the same time, the mass media have started using it more and more; in fact, there are in this country [Nigeria] 3 weekly newspapers published in this language. As for education, Hausa is optional at primary school level in Nigeria and obligatory at the secondary school level. In Niger, on the other hand, there is a pilot project to use Hausa as a language of instruction in secondary as well as in higher education.
>
> *Songhoy*
>
> This language is known as Zarma in Niger and Nigeria. Dendi, which is spoken in Benin, is considered as the same language as Songhoy because of mutual intelligibility.
>
> Songhoy is the second language in Niger, in terms of the number of speakers ... in Mali it is the third. A lot of research on this language is being conducted in various countries ... in order to use it as a means to undertake literacy programmes in these areas. As for education, Songhoy is utilized as a language of instruction in the first three years of primary school. It is also taught at the university as an optional subject.
>
> This language is also used in the mass media with a monthly newspaper selling 500 copies in Benin and three other monthlies in Niger with a circulation of 3,000 copies for one and 1,000 copies each of the remaining two.
>
> *Wolof*
>
> This language is currently spoken in Senegal, the Gambia and Mauritania. It is considered in all three countries as a community language. ... In all three countries where it is used, there are several pilot projects to use it as a language of instruction at the primary level of education; ... As for literacy, there are 43,000 people in the Gambia who have been initiated to this language, which is written in the Arabic script. There is a similar activity in Senegal, where the authorities use television to teach people to write the language.

Chtatou then goes on to discuss the desire in several African countries to develop Arabic-script orthographies for these languages:

> ...in the wake of decolonization, the interest in these languages was revived and governments set out to give them the status the deserve on a national level. ... These languages were also used to promote much-needed literacy programmes for people of different ages. As for some countries of Muslim Africa, their top priority was to devise for their languages an Arabic script in which they can be written so that the language can be used for the purpose of promoting literacy...
>
> Keen on the development of such a script for their languages, African governments approached such international organizations as UNESCO and ISESCO ... The first workshops were organized by UNESCO through BREDA (Bureau Régional d'Education pour l'Afrique), and later on ISESCO joined in the effort and sponsored other workshops on the same topic.

The bulk of Chtatou's work consists of reports on the writing conventions adopted at these workshops during the late 1980s. The charts he shows of existing and proposed transcription systems (either already in use or proposed as a result of the effort to standardize the writing systems) include a number of Arabic-script letters that are not supported in Unicode. This proposal, therefore, aims to extend the UCS repertoire to include the African characters documented in this study, as well as additional characters found in other African-language publications (see References).

3.1.1 Base (consonant) characters

The following 20 extended Arabic letters are found in the sources for African languages. All these proposed characters are of General Category Lo; Combining Class 0; Bidi Type AL. The suggested codepoints are those approved by the UTC in June 2003. (Gaps in the sequence of codepoints will be filled with characters discussed in subsequent sections of this document.) These characters are placed in a proposed new *Arabic Supplementary* block at U+0750, leaving the remaining spaces in the U+0600 block for Arabic-script diacritics, punctuation, etc.

Glyph	*Code*	*Character name*	*Shaping*	*See figures*
ﭖ	0750	ARABIC LETTER BEH WITH THREE DOTS HORIZONTALLY BELOW	BEH	3, 7, 9
ﭡ	0751	ARABIC LETTER BEH WITH DOT BELOW AND THREE DOTS ABOVE	BEH	4, 7, 9
ﭢ	0752	ARABIC LETTER BEH WITH THREE DOTS POINTING UPWARDS BELOW	BEH	4, 7, 8, 10, 11
ﭣ	0753	ARABIC LETTER BEH WITH THREE DOTS POINTING UPWARDS BELOW AND TWO DOTS ABOVE	BEH	5
ﭤ	0754	ARABIC LETTER BEH WITH TWO DOTS BELOW AND DOT ABOVE	BEH	7, 8, 9
ﭥ	0755	ARABIC LETTER BEH WITH INVERTED SMALL V BELOW	BEH	10
ﭦ	0756	ARABIC LETTER BEH WITH SMALL V	BEH	10
ﭧ	0757	ARABIC LETTER HAH WITH TWO DOTS ABOVE	HAH	2, 3, 7, 8
ﭨ	0758	ARABIC LETTER HAH WITH THREE DOTS POINTING UPWARDS BELOW	HAH	8
ﭩ	075A	ARABIC LETTER DAL WITH INVERTED SMALL V BELOW	DAL	11
ﭪ	075B	ARABIC LETTER REH WITH STROKE THROUGH	REH	11
ﭫ	075D	ARABIC LETTER AIN WITH TWO DOTS ABOVE	AIN	2, 3, 7, 9, 10, 11, 12, 13
ﭬ	075E	ARABIC LETTER AIN WITH THREE DOTS POINTING DOWNWARDS ABOVE	AIN	6, 8
ﭭ	075F	ARABIC LETTER AIN WITH TWO DOTS VERTICALLY ABOVE	AIN	8

ڤ	0760	ARABIC LETTER FEH WITH TWO DOTS BELOW	FEH	3, 7, 8, 9
ڥ	0761	ARABIC LETTER FEH WITH THREE DOTS POINTING UPWARDS BELOW	FEH	11
ڿ	0764	ARABIC LETTER KEHEH WITH THREE DOTS POINTING UPWARDS BELOW	GAF	11
۾	0765	ARABIC LETTER MEEM WITH DOT ABOVE	MEEM	3, 7
ۿ	0766	ARABIC LETTER MEEM WITH DOT BELOW	MEEM	12, 13
ݨ	0767	ARABIC LETTER NOON WITH TWO DOTS BELOW AND DOT ABOVE	NOON	1, 11, 13

3.1.2 Vowel signs

Several new signs have been used in African languages to represent vowel sounds not present in standard Arabic. The following three signs are proposed for encoding as combining characters in the Arabic block. The codepoints shown here are are those approved by the UTC in June 2003. (Note that the ARABIC ZWARAKAY, as proposed in L2/03-133 (Everson & Pournader 2003), is to be encoded at U+0659.)

Glyph	*Code*	*Character name*	*GC*	*CC*	*Bidi*	*See figures*
◌ٚ	065A	ARABIC VOWEL SIGN SMALL V ABOVE	Mn	30	NSM	12, 13
◌ٛ	065B	ARABIC VOWEL SIGN INVERTED SMALL V ABOVE	Mn	30	NSM	12, 13
◌ٜ	065C	ARABIC VOWEL SIGN DOT BELOW	Mn	32	NSM	14, 15, 16, 17

3.1.3 Samples showing African characters

The consonant and vowel characters listed above are illustrated in the following samples from Chtatou (1992) and other sources.

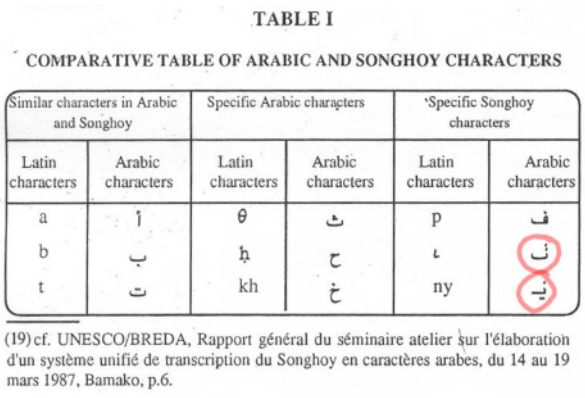


TABLE I

COMPARATIVE TABLE OF ARABIC AND SONGHOY CHARACTERS

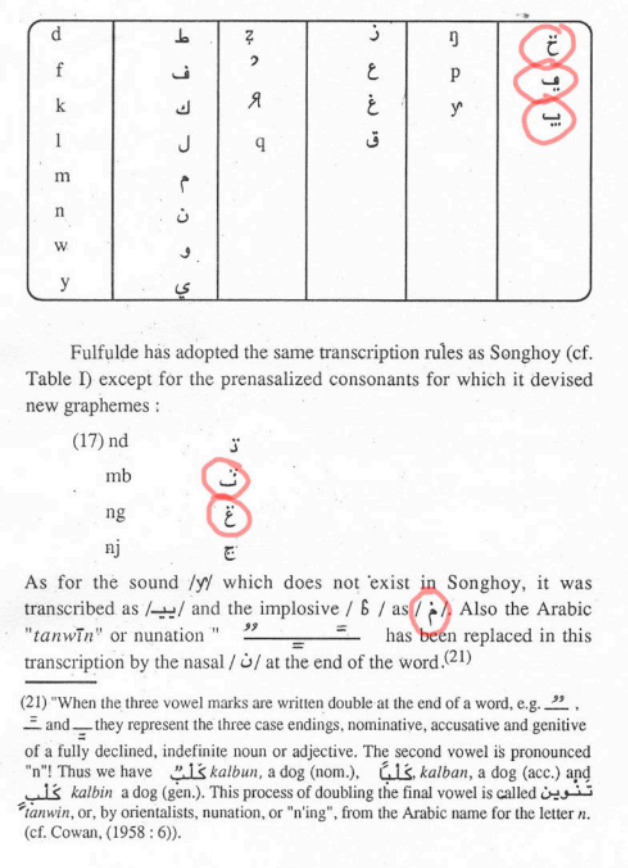
Similar characters in Arabic and Songhoy		Specific Arabic characters		Specific Songhoy characters	
Latin characters	Arabic characters	Latin characters	Arabic characters	Latin characters	Arabic characters
a	أ	θ	ث	p	ف
b	ب	ḥ	ح	ɛ	ڤ
t	ت	kh	خ	ny	ڹ

(19) cf. UNESCO/BREDA, Rapport général du séminaire atelier sur l'élaboration d'un système unifié de transcription du Songhoy en caractères arabes, du 14 au 19 mars 1987, Bamako, p.6.

Figure 1: Chtatou (1992), page 28. This also shows a BEH WITH TWO DOTS VERTICALLY ABOVE RIGHT SIDE, but it is unclear whether this needs to be distinguished from U+067A, and it is therefore not proposed for encoding at this time.

j	ج	ʒ	ز	ŋ	ڠ
d	د	ɛ	ص	g	غ
r	ر	ḍ	ض		
z	ز	ṭ	ط		
s	س	ǯ	ظ		
š	ش	ɾ	ع		
f	ف	ʁ	غ		
k	ك	q	ق		
l	ل				
m	م				
n	ن				
h	ه				
w	و				
y	ي				

Figure 2: Chtatou (1992), page 29.



d	ط	ʐ	ز	ŋ	خ
f	ف	ʔ	ع	p	ڢ
k	ك	ʀ	غ	ɣ	ڢ
l	ل	q	ق		
m	م				
n	ن				
w	و				
y	ي				

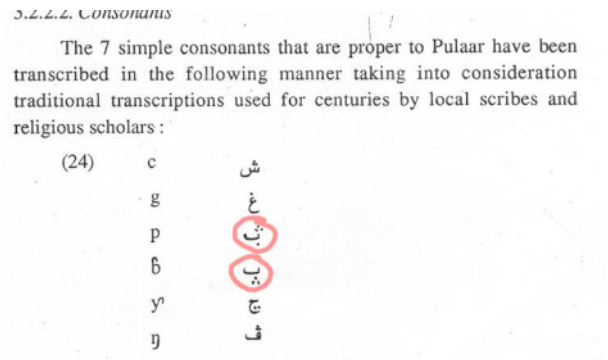
Fulfulde has adopted the same transcription rules as Songhoy (cf. Table I) except for the prenasalized consonants for which it devised new graphemes :

(17) nd ت
mb ٮ
ng غ
nj ج

As for the sound /ɣ/ which does not exist in Songhoy, it was transcribed as /ـيـ/ and the implosive / ɓ / as / م /. Also the Arabic "*tanwīn*" or nunation " ـٌ ـٍ ـً has been replaced in this transcription by the nasal / ن / at the end of the word.(21)

(21) "When the three vowel marks are written double at the end of a word, e.g. ـٌ , ـً and ـٍ they represent the three case endings, nominative, accusative and genitive of a fully declined, indefinite noun or adjective. The second vowel is pronounced "n"! Thus we have كَلْبٌ *kalbun*, a dog (nom.), كَلْبًا *kalban*, a dog (acc.) and كَلْبٍ *kalbin* a dog (gen.). This process of doubling the final vowel is called تَنْوِين *tanwin*, or, by orientalists, nunation, or "n'ing", from the Arabic name for the letter *n*. (cf. Cowan, (1958 : 6)).

Figure 3: Chtatou (1992), page 34. This also shows a BEH WITH THREE DOTS POINTING DOWNWARDS ABOVE RIGHT SIDE, but it is unclear whether this needs to be distinguished from U+067D, and it is therefore not proposed for encoding at this time.

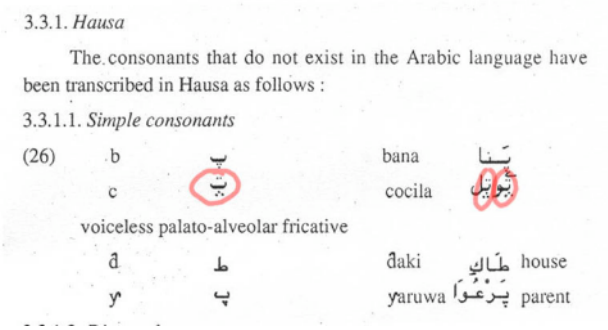


3.2.2.2. Consonants

The 7 simple consonants that are proper to Pulaar have been transcribed in the following manner taking into consideration traditional transcriptions used for centuries by local scribes and religious scholars :

(24) c ش
g غ
p ٮ
ɓ ڢ
ɣ ج
ŋ ڤ

Figure 4: Chtatou (1992), page 39. The positioning of the dots over/under the right end of the base form, rather than centrally, is considered to be an idiosyncrasy that does not merit separate encoding.



3.3.1. *Hausa*

The consonants that do not exist in the Arabic language have been transcribed in Hausa as follows :

3.3.1.1. *Simple consonants*

(26) b پ bana بَنا

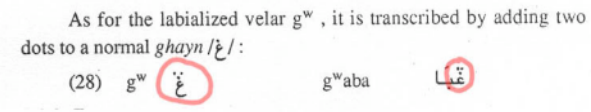
c ڀ cocila ڀُوڀِل

voiceless palato-alveolar fricative

ɗ ط ɗaki طَاكِ house

ɣ پ ɣaruwa پَرْعُوَا parent

Figure 5: Chtatou (1992), page 42.



As for the labialized velar g^w , it is transcribed by adding two dots to a normal *ghayn* /غ/ :

(28) g^w غٛ g^waba غٛبَا

Figure 6: Chtatou (1992), page 43.

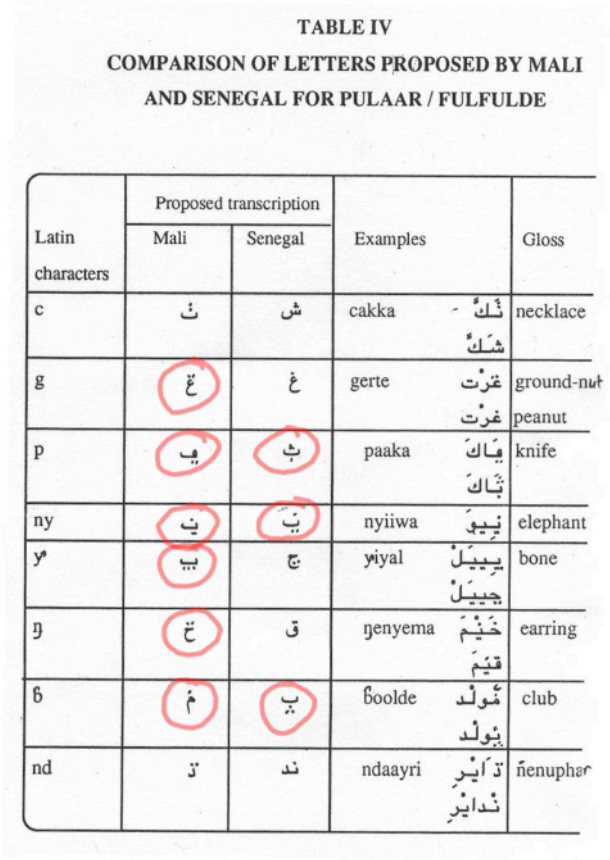


TABLE IV

COMPARISON OF LETTERS PROPOSED BY MALI AND SENEGAL FOR PULAAR / FULFULDE

Latin characters	Proposed transcription: Mali	Proposed transcription: Senegal	Examples		Gloss
c	ٺ	ش	cakka	ٺَكٌ شَكٌ	necklace
g	ڠ	غ	gerte	ڠرْت غرْت	ground-nut peanut
p	ڢ	ڢ	paaka	ڢَاكَ ڢَاكَ	knife
ny	ڽ	ڽ	nyiiwa	ڽِيوَ	elephant
ɣ	ڽ	ڽ	ɣiyal	ڽِييَلْ ڽِييَلْ	bone
ŋ	ڭ	ق	ŋenyema	ڭَيْمَ قيْمَ	earring
ɓ	ڢ	ڢ	ɓoolde	ڢُولْد ڢُولْد	club
nd	ڌ	ند	ndaayri	ڌَايْرِ نْدايْرِ	ñenuphar

Figure 7: Chtatou (1992), page 45. It seems clear that the writer is making a conscious distinction between three dots in a horizontal row and three dots in a triangle formation here.

TABLE V

COMPARISON OF LETTERS PROPOSED BY MALI AND NIGER FOR THE TRANSCRIPTION OF ZARMA / SONGHOY

Latin characters	Proposed transcription: Mali	Proposed transcription: Niger	Examples	Gloss	
c	ٺ	ڇ	ciiri	نِيرُ چِير	salt
g	ڠ	غ	gaara	ڠَارَ غَارَ	to sollicit
ny	ڽ	ٽ	nyaamoy	ڽَامَىْ ٽَامَىْ	
p	ڢ	ڀ	paate	ڢَات ڀَات	
ŋ	ڭ	ڠ	naari	ڭَارِ ڠَارِ	rice
o	ـَ	ـَ	koyra	كَيْرَ	village

Figure 8: Chtatou (1992), page 47.

TABLE VII

PULAAR / FULFULDE STANDARDIZED ARABIC ALPHABET

Similar characters in Arabic and Pulaar/Fulfulde	Specific characters to Pulaar/Fulfulde	Pulaar/Fulfulde prenasalized sounds	Pulaar/Fulfulde vowels
أ	ٺ	مب	ـَ ـِ ـُ
ب	ڀ	ند	ـٰ ـٰ ـٰ
ت	ٿ	نج	
ج	ڠ	نع	
د	ڃ		
ر	ڽ		
س	ڢ		
ط			
ف			
ك			
ل			
م			
ن			
هـ			
و			
ي			
ق			

Figure 9: Chtatou (1992), page 50.

(34)

ا without *hamza* for **a, e, i, o, u**

ٱ without *fatḥa* for **aa**

پ Pesian and Urdu symbol for **p** (instead of ڀ).

ڀ for **ɓ** (instead of ب)

ٺ for **c** (instead of ث)

ج for **y'** (no change)

ط for **ɗ** (no change)

گ Persian and Urdu symbol for **g** (instead of غ)

ڠ for **ŋ** (instead of ڤ)

ـَ for the vowel **e**

Figure 10: Chtatou (1992), page 54.

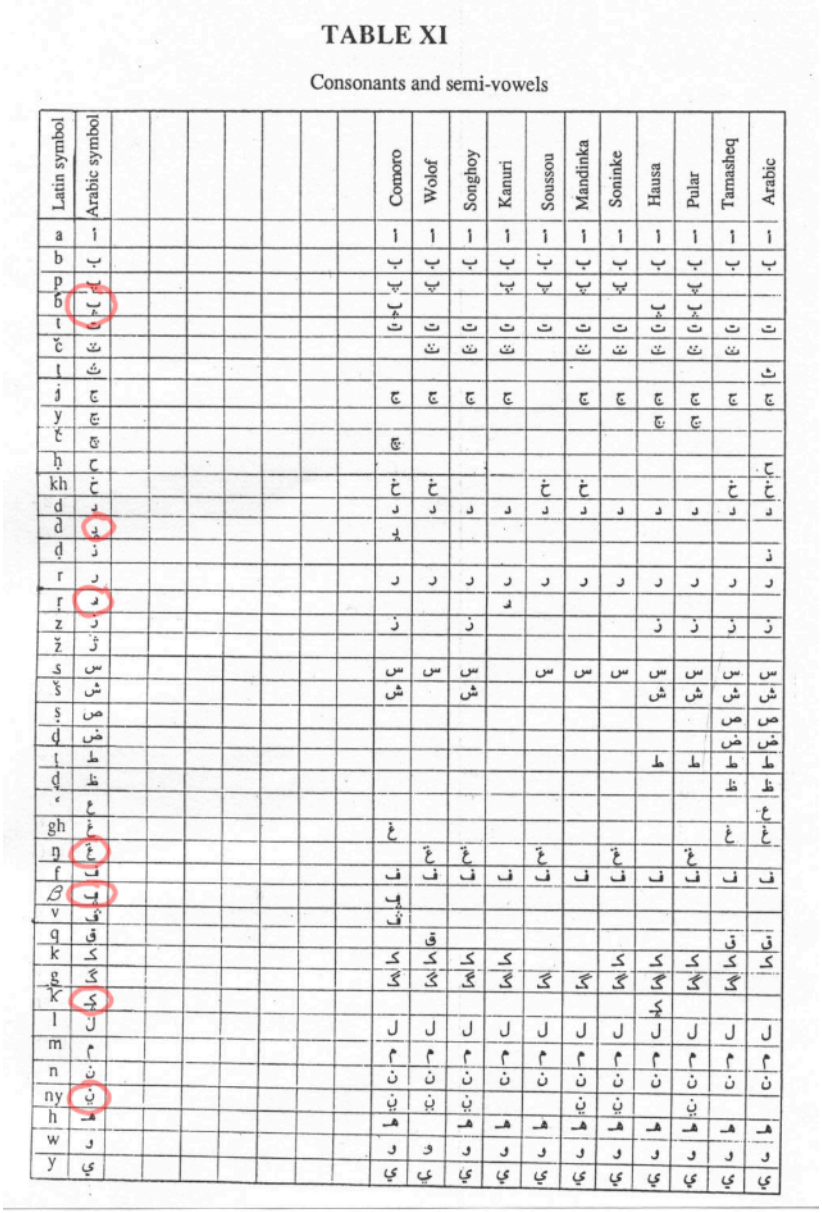


TABLE XI

Consonants and semi-vowels

Latin symbol	Arabic symbol	Comoro	Wolof	Songhoy	Kanuri	Soussou	Mandinka	Soninke	Hausa	Pular	Tamasheq	Arabic
a	ا	ا	ا	ا	ا	ا	ا	ا	ا	ا	ا	ا
b	ب	ب	ب	ب	ب	ب	ب	ب	ب	ب	ب	ب
p	پ	پ	پ		پ	پ	پ	پ		پ		
b	ɓ	ɓ	ɓ						ɓ	ɓ		
t	ت	ت	ت	ت	ت	ت	ت	ت	ت	ت	ت	ت
c	ث		ث	ث								
j	ج	ج	ج	ج	ج		ج	ج	ج	ج	ج	ج
y	ڃ								ڃ	ڃ		
č	چ	چ										
ḥ	ح											ح
kh	خ	خ	خ			خ	خ				خ	خ
d	د	د	د	د	د	د	د	د	د	د	د	د
ḍ	ڈ											
ḍ	ض	ض										ض
r	ر	ر	ر	ر	ر	ر	ر	ر	ر	ر	ر	ر
ṛ	ڑ			ڑ	ڑ							
z	ز	ز		ز		ز			ز	ز	ز	ز
ž	ژ											
s	س	س	س	س		س	س	س	س	س	س	س
š	ش	ش		ش					ش	ش	ش	ش
ṣ	ص	ص									ص	ص
ḍ	ض	ض									ض	ض
ṭ	ط	ط							ط	ط	ط	ط
ḍ	ظ	ظ									ظ	ظ
ʿ	ع	ع										ع
gh	غ	غ									غ	غ
ŋ	ڭ		ڭ	ڭ		ڭ		ڭ		ڭ		
f	ف	ف	ف	ف	ف	ف	ف	ف	ف	ف	ف	ف
β	ڤ	ڤ										
v		ڤ										
q	ق		ق	ق							ق	ق
k	ك	ك	ك	ك	ك	ك	ك	ك	ك	ك	ك	ك
g	گ	گ	گ	گ	گ	گ	گ	گ	گ	گ	گ	
k	ک	ک							ک			
l	ل	ل	ل	ل	ل	ل	ل	ل	ل	ل	ل	ل
m	م	م	م	م	م	م	م	م	م	م	م	م
n	ن	ن	ن	ن	ن	ن	ن	ن	ن	ن	ن	ن
ny	ڽ	ڽ	ڽ	ڽ			ڽ	ڽ		ڽ		
h	ه	ه		ه	ه	ه	ه	ه	ه	ه	ه	ه
w	و	و	و	و	و	و	و	و	و	و	و	و
y	ي	ي	ي	ي	ي	ي	ي	ي	ي	ي	ي	ي

Figure 11: Chtatou (1992), page 58. Note that most of the characters shown in this chart are already encoded; only the indicated ones are relevant to this proposal.

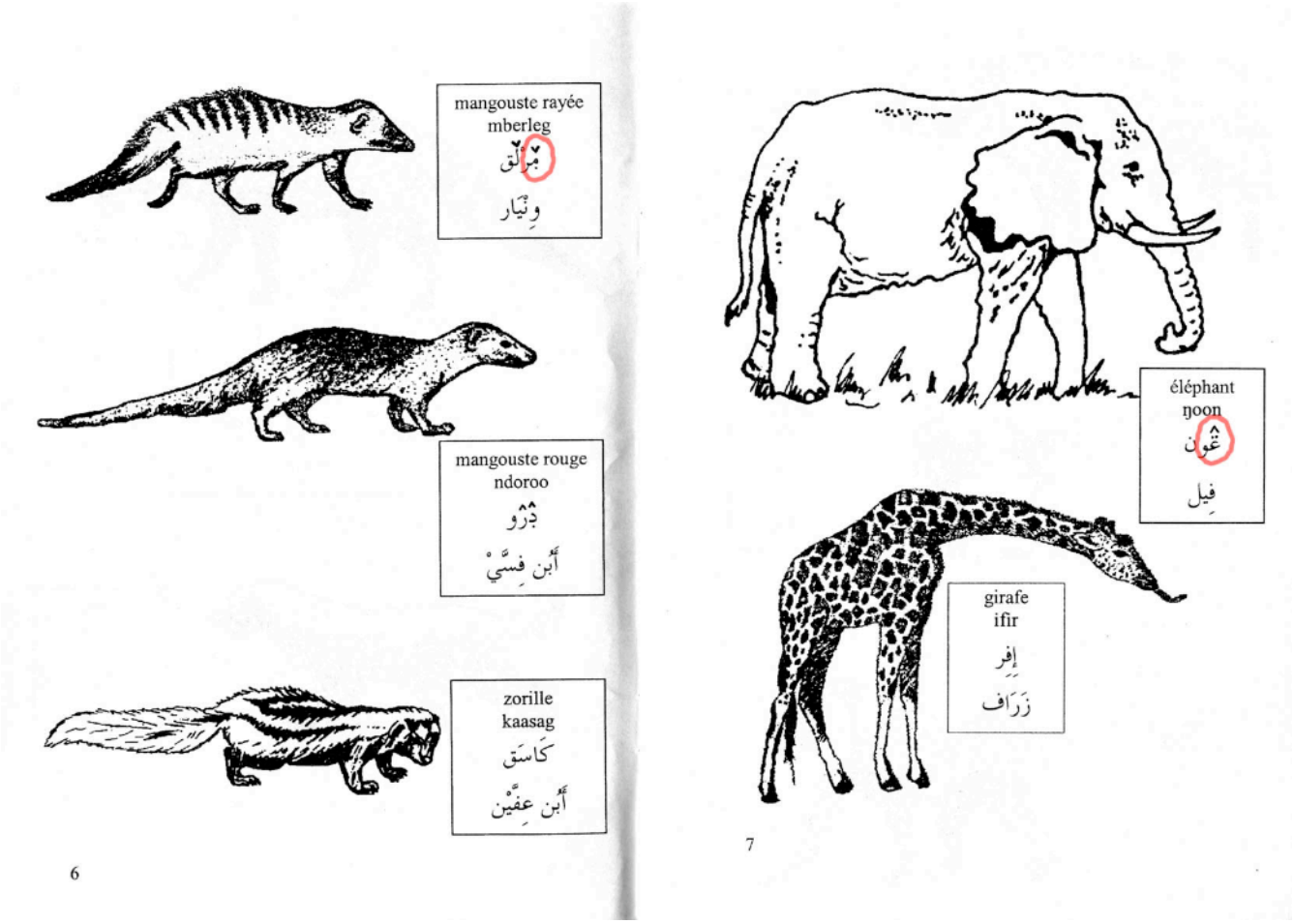
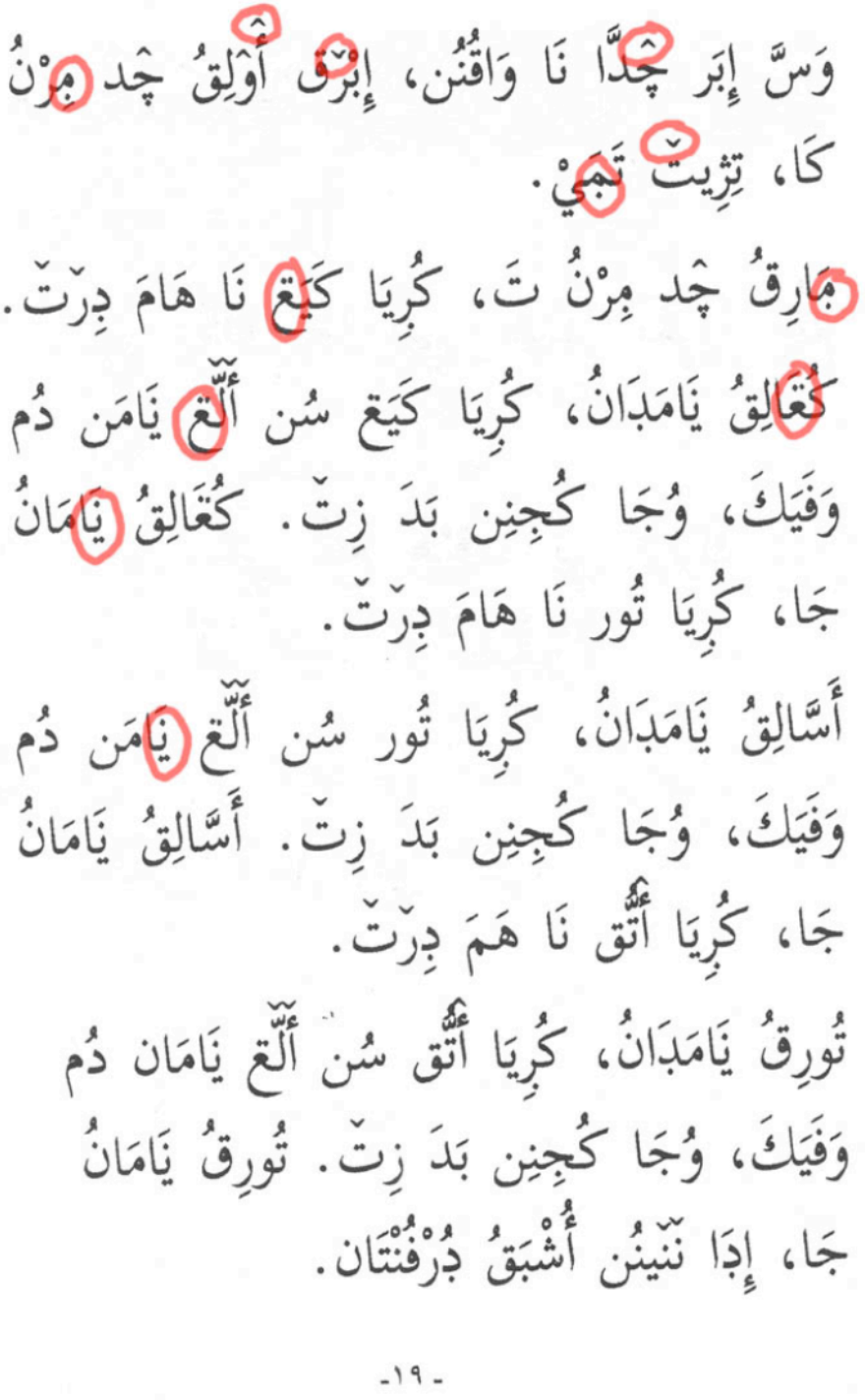


Figure 12: Nodjindaina et al (2002), pages 6–7.



وَسَّ إِبَر چَدَّا نَا وَاقُنُن، إِبْرَق أَوْلِقُ چَد مِرْنُ
كَا، تِزِيتّ تَمِيْ.

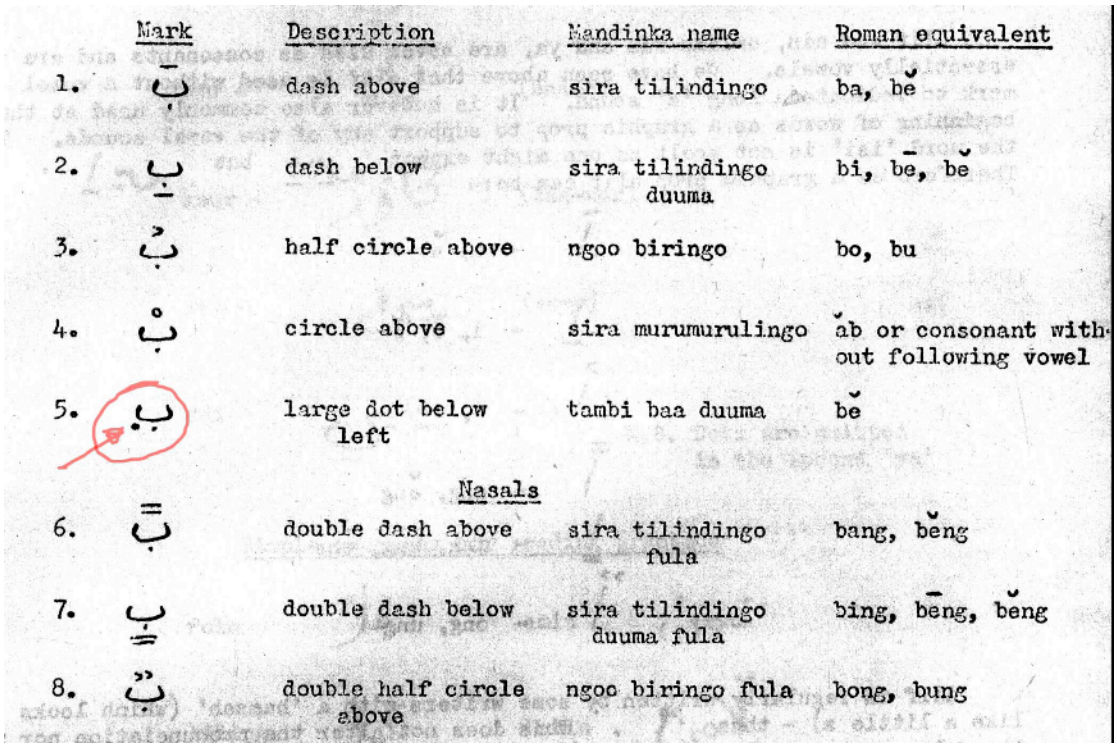
مَارِقُ چَد مِرْنُ تَ، كُرِيَا كَيَع نَا هَامَ دِرّتّ.

كُتَعَالِقُ يَامَدَانُ، كُرِيَا كَيَع سُن أَلّغ يَامَن دُم
وَفَيَكَ، وُجَا كُجِنِن بَدَ زِتّ. كُتَعَالِقُ يَامَانُ
جَا، كُرِيَا تُور نَا هَامَ دِرّتّ.

أَسَّالِقُ يَامَدَانُ، كُرِيَا تُور سُن أَلّغ يَامَن دُم
وَفَيَكَ، وُجَا كُجِنِن بَدَ زِتّ. أَسَّالِقُ يَامَانُ
جَا، كُرِيَا أُتّق نَا هَمَ دِرّتّ.

تُورِقُ يَامَدَانُ، كُرِيَا أُتّق سُن أَلّغ يَامَان دُم
وَفَيَكَ، وُجَا كُجِنِن بَدَ زِتّ. تُورِقُ يَامَانُ
جَا، إِدَا نّنِنُن أُشْبَقُ دُرْفُنْتَان.

Figure 13: Dahab et al (2002), page 19.

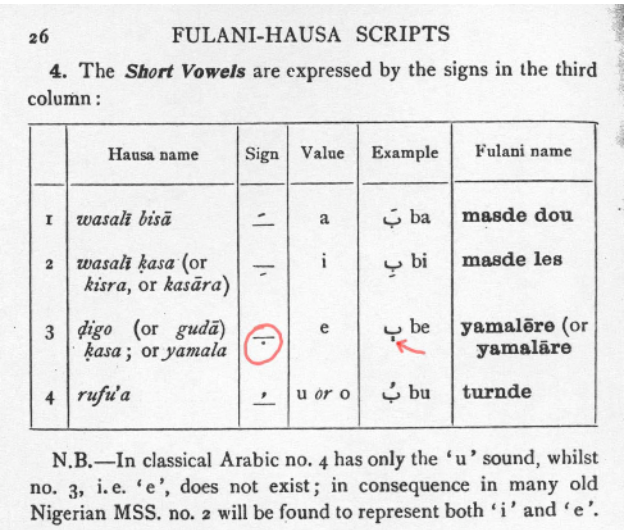


	Mark	Description	Mandinka name	Roman equivalent
1.	بَ	dash above	sira tilindingo	ba, bĕ
2.	بِ	dash below	sira tilindingo duuma	bi, bē, bĕ
3.	بُ	half circle above	ngoo biringo	bo, bu
4.	بْ	circle above	sira murumurulingo	ăb or consonant without following vowel
5.	بٖ	large dot below left	tambi baa duuma	bĕ
		Nasals		
6.	بّ	double dash above	sira tilindingo fula	bang, bĕng
7.	بٍ	double dash below	sira tilindingo duuma fula	bing, bēng, bĕng
8.	بٌ	double half circle above	ngoo biringo fula	bong, bung

Figure 14: Addis (1963), page 9: see item 5 in list of vowel marks.

	ɓ			بٖ	
c	c		ش	ش	ش
d	d	د			د
	ɗ			ط	ط
e	e		ـٖـ	ـٖـ	ـٖـ
é			ـٖـ		ـٖـ
ë			ـٖـ		ـٖـ
f	f	ف			ف

Figure 15: Chtatou (1992), page 40: see transcriptions for /e/. Results of *Workshop on the formulation of a standardized system of transcription of Wolof and Pulaar, 16–21 March 1987, Dakar, Senegal.*



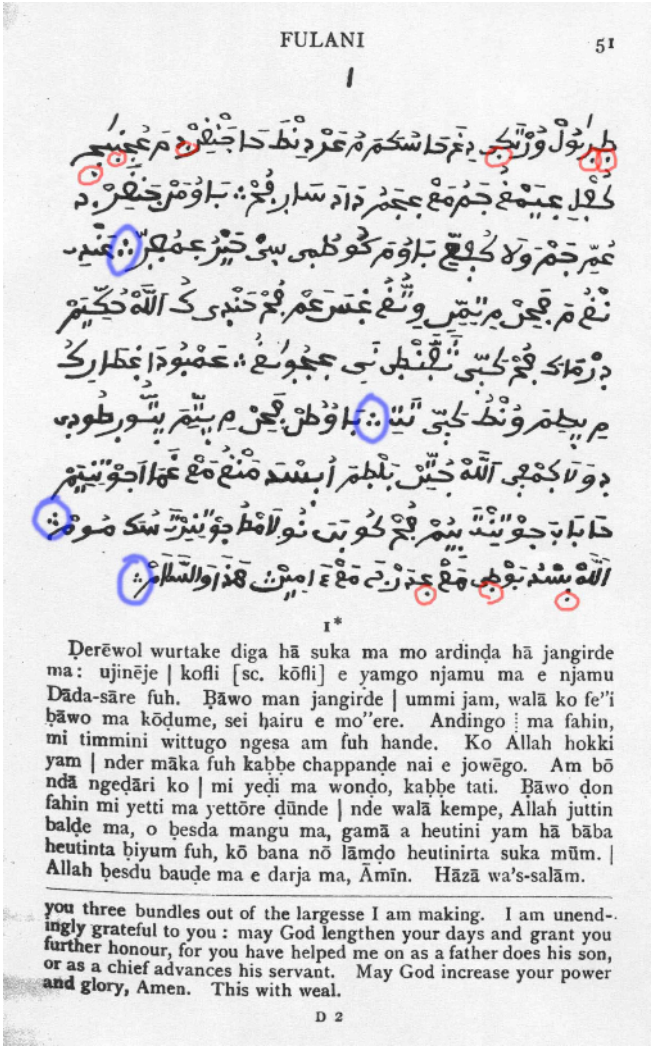
26 FULANI-HAUSA SCRIPTS

4. The ***Short Vowels*** are expressed by the signs in the third column:

	Hausa name	Sign	Value	Example	Fulani name
1	*wasalī bisā*	ـَ	a	بَ ba	**masde dou**
2	*wasalī ḳasa* (or *kisra*, or *kasāra*)	ـِ	i	بِ bi	**masde les**
3	*ḍigo* (or *gudā*) *ḳasa*; or *yamala*	ـِ·	e	بِ be	**yamalēre** (or **yamalāre**
4	*rufu'a*	ـُ	u *or* o	بُ bu	**turnde**

N.B.—In classical Arabic no. 4 has only the 'u' sound, whilst no. 3, i.e. 'e', does not exist; in consequence in many old Nigerian MSS. no. 2 will be found to represent both 'i' and 'e'.

Figure 16: Taylor (1929), page 26: see item 3 in table of vowel signs.



FULANI 51

1

1*

Ḍerēwol wurtake diga hā suka ma mo ardinḍa hā jangirde ma: ujinēje | kofli [sc. kōfli] e yamgo njamu ma e njamu Dāda-sāre fuh. Ḅāwo man jangirde | ummi jam, walā ko fe"i ḅāwo ma kōḍume, sei ḥairu e mo"ere. Andingo | ma fahin, mi timmini wittugo ngesa am fuh hande. Ko Allah hokki yam | nder māka fuh kabḅe chappanḍe nai e jowēgo. Am bō ndā ngeḍāri ko | mi yeḍi ma wonḍo, kabḅe tati. Ḅāwo ḍon fahin mi yetti ma yettōre ḍūnde | nde walā kempe, Allah juttin balḍe ma, o ḅesda mangu ma, gamā a heutini yam hā bāba heutinta ḅiyum fuh, kō bana nō lāmḍo heutinirta suka mūm. | Allah ḅesdu bauḍe ma e darja ma, Āmīn. Hāzā wa's-salām.

you three bundles out of the largesse I am making. I am unendingly grateful to you: may God lengthen your days and grant you further honour, for you have helped me on as a father does his son, or as a chief advances his servant. May God increase your power and glory, Amen. This with weal.

D 2

Figure 17: Taylor (1929), page 51: showing use of dot for /e/ in running text (Fulani), and triple-dot punctuation mark (see section 3.4).

3.2 Jawi and Moroccan GAF characters

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 18:

ك ک ڪ

A B C

Figure 18: 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 19:

qāf	q, ḳ	[q]	100	ق	ق	ق	ڤ
kāf	k	[k]	20	ك	ك	ک	ڪ
lām	l	[l]	30	ل	ل	ل	ڸ

Figure 19: 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 20; the typical Arabic form A (ك) is not used in Sindhi.

q	[k]	ق	ق	ق	ڤ
k	[k]	ڪ	ڪ	ڪ	ڪ
kh	[k^h]	ک	ک	ک	ک
g	[g]	گ	گ	گ	گ

Figure 20: 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 21: 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 21 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*.

The following two characters, representing *gaf* as written in different regions, are therefore proposed for encoding; they are discussed individually following the summary table. Both share similar Unicode properties: general category Lo; combining class 0; bidi type AL.

Glyph	*Code*	*Character name*	*Shaping*	*See figures*
ک	0762	ARABIC LETTER KEHEH WITH DOT ABOVE	GAF	22, 24
ڭ	0763	ARABIC LETTER KEHEH WITH THREE DOTS ABOVE	GAF	25, 26, 27, 28, 29

3.2.1 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	ڤ	ڤ
k	[k]	ك	ك	ک	ک
g	[g]	ک	ک	ک	ک
l	[l]	ل	ل	ل	ل

Figure 22: From chart of the Jawi alphabet, Daniels & Bright (1996), page 761.

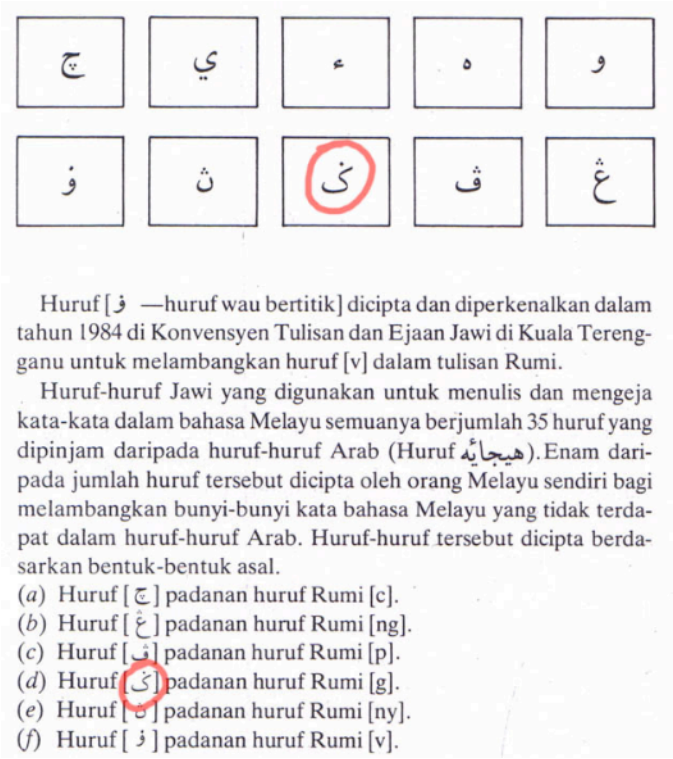
Note in figure 22 that form B is used as the basis for the /g/ character, despite the fact that form A is used for /k/. This is a characteristic of Jawi writing, and not an artifact of this particular book's typography. Comparing the chart for Uighur, found on the previous page, we see that in some cases 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	ڤ	ڤ
k	[k]	ك	ك	ک	ک
g	[g]	گ	گ	گ	گ
ng	[ŋ, n]	ڭ	ڭ	ڭ	ڭ
l	[l]	ل	ل	ل	ل

Figure 23: From chart of the Uighur alphabet, Daniels & Bright (1996), page 760.

Figure 23 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 24 is taken from an introduction to the Jawi script published in Malaysia.


چ ي ء ه و

ۏ ث ک ڤ غ

Huruf [ۏ —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 [ڤ] padanan huruf Rumi [p].
(*d*) Huruf [ک] padanan huruf Rumi [g].
(*e*) Huruf [ث] padanan huruf Rumi [ny].
(*f*) Huruf [ۏ] padanan huruf Rumi [v].

Figure 24: 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 ک 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 DOT ABOVE should be encoded, and a note added to U+06AC indicating that the new character is preferred for old Malay.

3.2.2 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 25: Street name in Arabic and Latin scripts.

Figure 26: Petrol pump, labeled 'gasoil' in Arabic script.

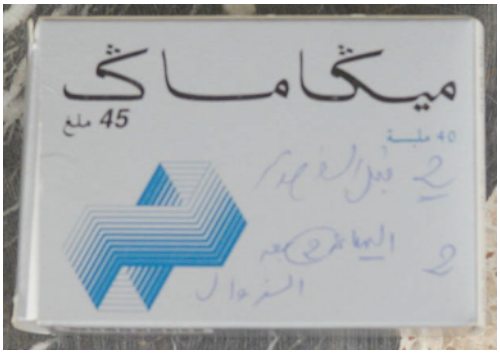


Figure 27: Medicine package 'Megamag', showing that the *gaf* is based on *kaf* form B.

Figure 28 is taken from the Royal Moroccan Academy's new Amazigh dictionary, and shows the letter in initial, medial, and isolated forms.

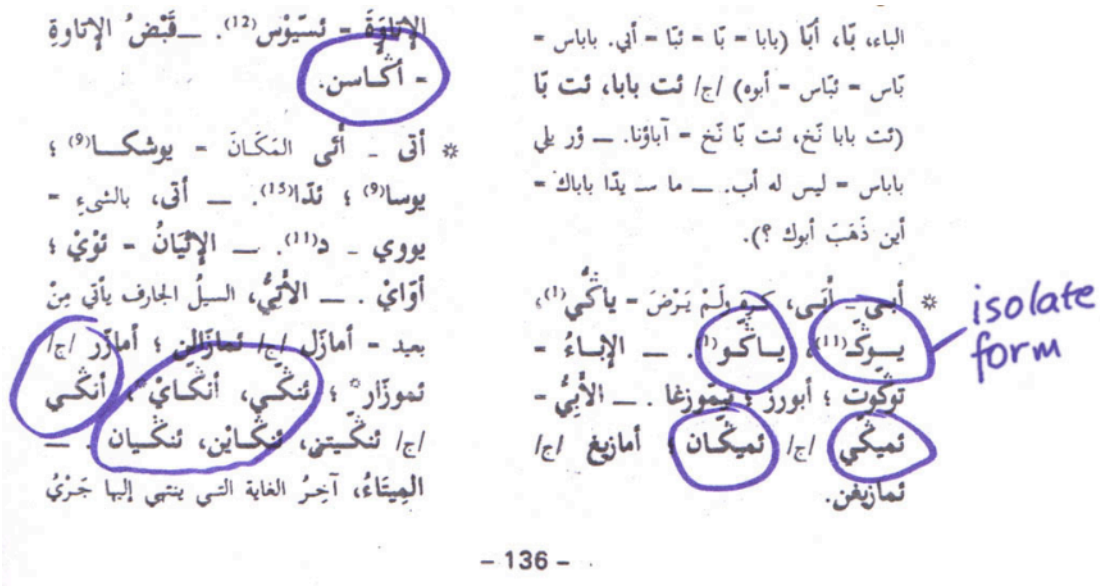


Figure 28: From Shafiq (1996), page 136; note the "makeshift" isolated *gaf*.

Close examination of figure 28 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* الگنيديّه	The 'white' way (*jaamba*) الطريق البيضاء
kar كر	[1]*entemaas* انتماس	*enweffal* انوقّل	[1]*mekka mūsa* مكّ موس
	[2]*sayni kar (ma yuḵarraṣ)* سيني كر (ما يخرّص)		[2]*el-faayez* الفايز
faaġu فاغ	[1]*tenaččuuga* تنچّوگه	*šš̆baar* 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 hayba*[3] سنّيمة هيب	('black') *el-mawsṭi* الموسطي	[1]*et-teḥzaam*[4] التّحزام
	[2]*meqaččuuga* (Hawd) مقچّوگه		[2]*eññaama*[5] انّيامه
	or *lebyaaḍ* لبياظ	('white') *menčalla* منچّله	[2]*rrbaabi*[6] الرّبابي
	or *etbaybi* (Trārza) اتبيبي		(a) *leggetri*[6] لگّتري
			(b) *čaynna* چينّه
			(c) *liyyin* ليّن
baygi or *lebtayt* بيگي	*baygi* بيگي	*el-muḵaalef* المخالف	*el-'itiig* العتيگ
	intensive form *a'ḍḍaal* اعضّال	or *baygi jjraad* بيگي الجّراد	
		or *baygi lekbiir* بيگي الكبير	

[1] Entry (*dḵuul* 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 *īggāwen*.
[5] Not subsidiary to those who include *et-teḥzaam* in *faaġu*.
[6] A blending of white and black.

Figure 29: 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 29).

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 users' minds; the fact that the forms ك and ڭ 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*.

3.3 South Asian extensions to Arabic script

The characters listed in this section have been used in writing several languages in the South Asia region; examples of the use of each character in one or more languages are included, but these should not be interpreted as representing the entire scope of use. It also seems likely that as additional minority languages in the same areas establish orthographic conventions, they may adopt some of these characters.

Four characters from South Asian languages are proposed for encoding; they are discussed further in individual subsections following the summary table. All share similar Unicode properties: general category Lo; combining class 0; bidi type AL.

Glyph	*Code*	*Character name*	*Shaping*	*See figures*
ڈ٘	0759	ARABIC LETTER DAL WITH TWO DOTS VERTICALLY BELOW AND SMALL TAH	DAL	32, 33, 34
ڜ	075C	ARABIC LETTER SEEN WITH FOUR DOTS ABOVE	SEEN	30, 31
ٹ٘ن	0768	ARABIC LETTER NOON WITH DOT ABOVE AND SMALL TAH	NOON	32, 33, 34, 35, 36
ں	0769	ARABIC LETTER NOON WITH SMALL V	NOON	37, 38

3.3.1 Retroflex DAL with two dots below

This character is used by Saraiki writers to represent an implosive retroflex /d/. Other Saraiki retroflexes are written with a small TAH above, following Urdu (e.g., U+0679, U+0688); other implosives are written with two dots vertically below, following Sindhi practice (e.g., U+067B, U+06B3). The natural tendency for a writer needing to write an implosive retroflex /d/, then, is to add the dots below a character like U+0688. Examples can readily be found in Saraiki newspapers and other books.

3.3.2 SEEN with four dots above

Writers of the Shina language in Kashmir, needing to write a retroflex /s/ letter, have adopted the convention of using a letter based on SHEEN but with four dots in place of three.

As a pattern of four dots is not normally seen in the area (though it does occur in southern Pakistani languages such as Sindhi; see U+067F, U+0680, etc.), some writers have tended to replace the four dots with two horizontal lines. This is seen in some of the examples (below). However, as it is common in handwritten script to see horizontal pairs of dots written as a single line, it seems appropriate to consider this a glyph or stylistic variation, encoding the character as SEEN WITH FOUR DOTS and leaving it to font designers to determine whether to offer glyphs with the dots replaced by lines. (It does appear from Taj (1989) that some writers specifically choose to write the retroflex /s/ with horizontal lines, even when well-formed dots are used on other letters. Nevertheless, it seems clear that the form with four dots and that with two lines are glyph variants of the same underlying character.)

It may be noted that Akbar (1985) also shows a character with the form of HAH WITH FOUR DOTS ABOVE used by a Shina writer (to represent the phoneme /ts/). The information currently available suggests that this

character has not been as widely adopted as the others shown here, with other writers using U+0685 or U+0697 for this sound, so the case for adding it to the UCS repertoire may be less clear-cut, at least until further data is obtained.

3.3.3 Retroflex NOON with dot

A number of languages use this character to represent a retroflex /n/. The use of a small TAH as part of an Arabic-script letter to indicate retroflexion is derived from Urdu usage, but Urdu itself does not have a retroflex /n/.

Sindhi has this sound, and writes it with a NOON where the dot is replaced by the small TAH (rather than the small TAH being added, while retaining the dot as well). However, this is not an option in languages where the orthography is based on Urdu, as the initial and medial forms would be indistinguishable from the retroflex /t/ (U+0679). Writers of such languages have therefore devised this letter, which is not yet encoded in Unicode.

To reduce the likelihood of confusion with the Sindhi retroflex /n/ letter (encoded as U+06BB ARABIC LETTER RNOON), the suggested name explicitly mentions the presence of the dot in this character; a name such as NOON WITH SMALL TAH might be taken to imply that the small TAH replaces the dot instead of being an addition, which would result in a character identical to U+06BB.

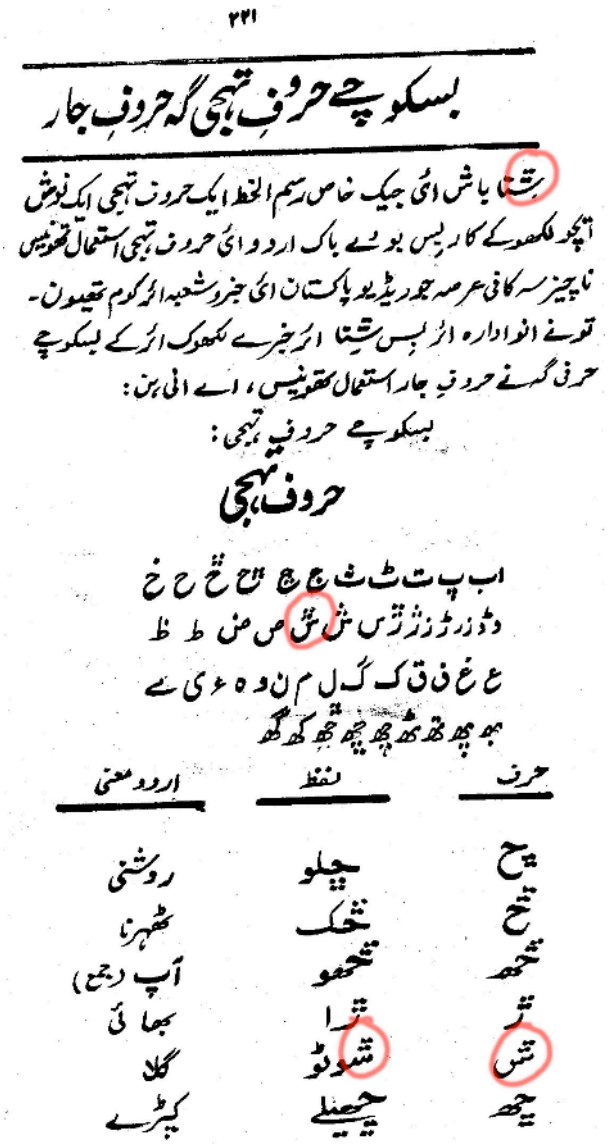
3.3.4 NOON with small V

The Gojri community is found in both India and Pakistan, and there has been a significant amount of literature published in both countries from the 1980s onwards. While there has been some variation in orthographic conventions, there is widespread use of a NOON WITH SMALL V to represent the retroflex nasal consonant. (Gojri writers also use a LAM WITH SMALL V, as seen in the examples, but this is already encoded at U+06B5.)

Phonologically, this character is used (at least in the Gojri language) to write the same sound as Sindhi RNOON or the NOON WITH DOT AND SMALL TAH proposed here, but it clearly represents a distinct choice of extended-Arabic character for this sound; the SMALL V and SMALL TAH show two different conventions for the creation of new Arabic-script letters. The fact that the phoneme being written may be the same is irrelevant to the encoding of the written characters.

3.3.5 Samples showing South Asian characters

The examples shown here are drawn largely from books and newspapers published in Pakistan (an Indian example is also included), written in Shina, Saraiki, Pathwari, and Gojri. However, it can be expected that as literacy becomes more widespread among minority language communities in South Asia, many of these writing conventions may be "borrowed" by neighboring communities. The languages cited here serve to demonstrate the need to encode these characters, but should not be assumed to be the only users of them.



۲۲۱

بسکوچے حروفِ تہجی گہ حروفِ جار

شِنا باش ائی جیک خاص رسم الخط ایک حروف تہجی ایک نوش
آیچو لکھوکے کار بِس بوڈے باک اردو ائی حروف تہجی استعمال تھونیس
ناچیز سہ کافی عرصہ جو ریڈیو پاکستان ائی خبرو شعبہ ائر کوم تھیون۔
تونے انو ادارہ ائر بِس شِنا ائر خبرے لکھوک ائر کے بسکوچے
حرفی گہ نے حروفِ جار استعمال تھونیس، اے انی ہن:
بسکوچے حروفِ تہجی:

حروف تہجی

اب پ ت ٹ ث ج چ ڃ ڄ ح خ
د ڈ ذ ر ڑ ز ژ ڙ س ش ڱ ص ض ط ظ
ع غ ف ق ک گ ل م ن و ہ ء ی ے
بھ پھ تھ ٹھ چھ چھ ڃھ کھ گھ

حرف	لفظ	اردو معنی
ڃ	ڃلو	روشنی
ڄ	ڄک	ٹھہرنا
ڄھ	ڄھو	آپ (جمع)
ڙ	ڙا	بھائی
ڱ	ڱوٹو	گلا
ڃھ	ڃھیلے	کپڑے

Figure 30: Akbar (1985), page 221 (Shina).

۲۹

Figure 31: Taj (1989), page 29 (Shina): examples of SEEN WITH FOUR DOTS ABOVE written using two lines in place of four dots (a practice also used for other four-dot letters in this book)

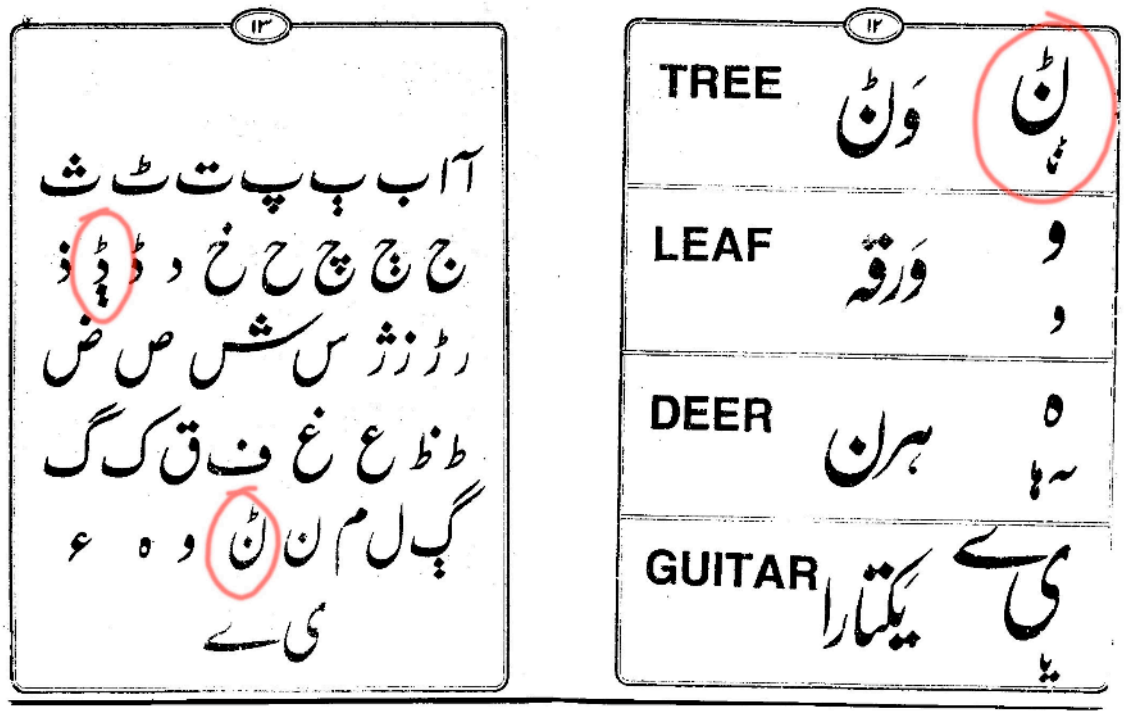


Figure 32: Mughal (1994), pages 12-13 (Saraiki).

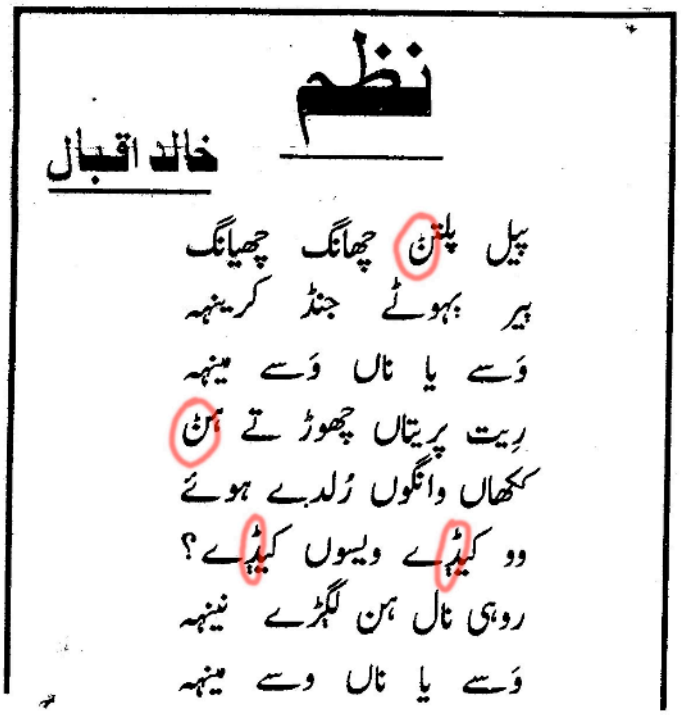


Figure 33: Saraiki poem printed in *Daily Jhoke*, 29 January 2002. The small size of the added TAH and TWO DOTS on the special Saraiki letters arises because these letters are not supported in Urdu software used to produce this newspaper, so these have been added manually. A similar size anomaly can be seen in the dots under BEEH (U+067B) in line 2 of the poem; this character is also not available in the Urdu font used, so the Urdu BEH (with one dot) has been typed, and a second dot added manually.

صرف

صرف لغت وچ بدلن یا ہیر پھیر کرن کوں آہدن ۔ علم وچ حروف تے انھاں دی حرکات (زبر ، زیر پیش) دے اول بدل نال انجو انج لفظ تے انھاں دے وکھرے وکھرے معنی حل آندن ۔ ایں گالھوں ایندا ناں علم صرف رکھیا گئے ۔

نحو

او علم ہے جیندے وچ کلام یعنی گالھ دے مختلف حصیں کوں صحیح طرحاں جوڑن تے بولن آندے تے انھاں حصیاں دے آپس دے تعلق دا پتہ لگدے ۔ تے انھاں غلطیاں کنوں بچیندے ۔ جنہاں دی وجہ نال کوئی گالھ سمجھ نہ آوے ، یا اوندا مطلب وگڑ ونجے ۔

حروف تہجی

کہیں زبان دیاں سادیاں اوازاں کوں حروف تہجی آکھیا ویندے ۔ جیویں جو ۔ ب ۔ پ ۔ ک ۔ م ۔ ل وغیرہ

سرائیکی دے مخصوص اضافی حروف تہجی

اردو دے سارے دے سارے حروف تہجی سرائیکی وچ وی استعمال کیتے ویندن البتہ خاص خاص اوازاں کیتے سرائیکی وچ پنج حروف اضافی طور تے شامل کیتے گئن ۔ او ایہ ہن ۔

ب ------ ہلی ۔ ہال ج ------ جالی ۔ جنج

ڈ ------ ڈانگ ۔ ڈولی گ ------ گاں ۔ لنگ

ٹ ------ وٹ ۔ پاٹی

Figure 34: Bhaya (1984), introductory page explaining special Saraiki letters

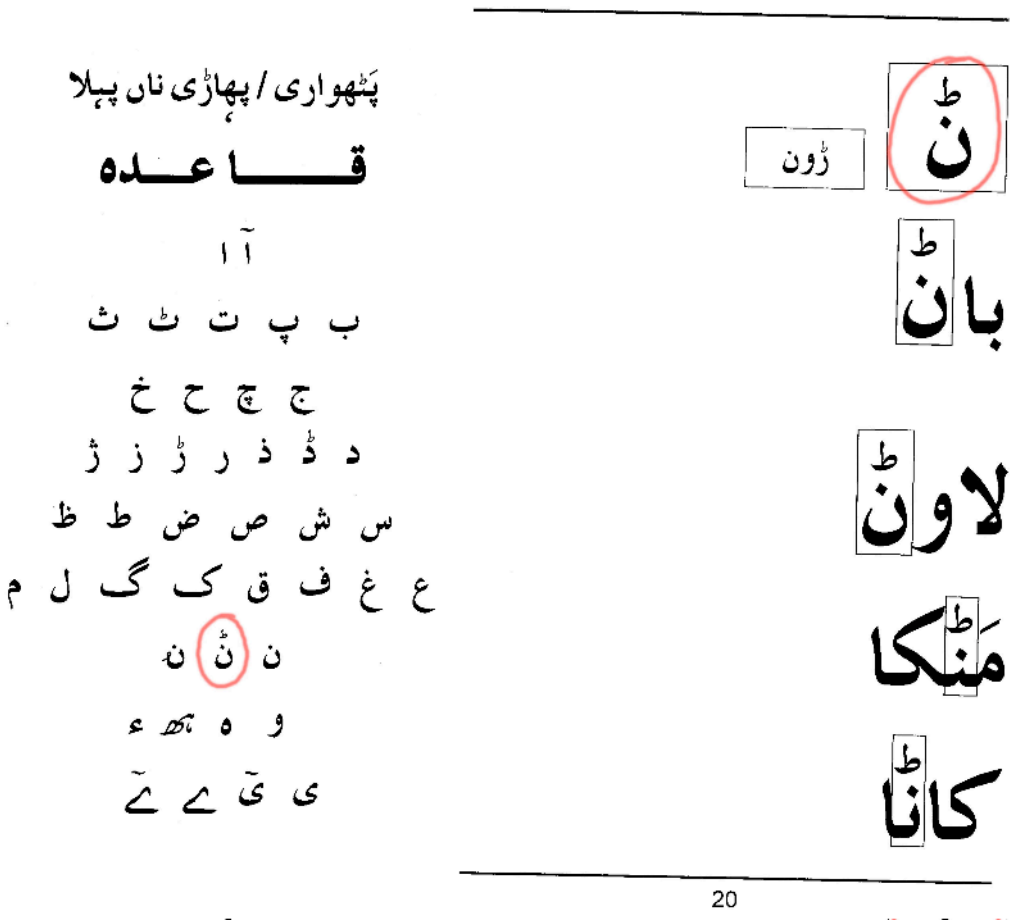


Figure 35: Chitka (n.d.), inside front cover and page 20 (Pathwari). The alphabet chart also shows a second "modified NOON" character, but it is unclear whether this form can be considered well enough established for standardization.

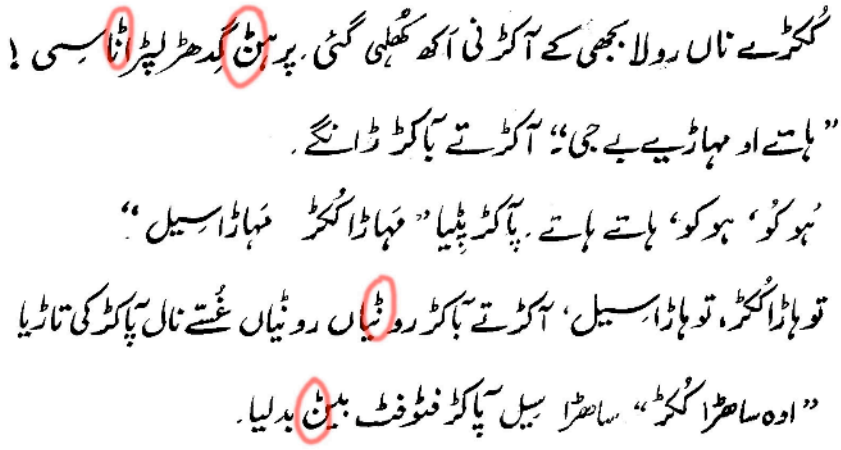

کُکڑے ناں رولا بجھی کے آکڑنی اَکھ کھُلہی گئی۔ پر ہن گِدھڑ لپڑا ناسی!
"ہاتے او مہاڑیے بے جی"۔ آکڑتے بآکڑ ڑانگے۔
'ہوکو' ہوکو' ہاتے ہاتے۔ پآکڑ پِٹیا "مہاڑا کُکڑ مہاڑا سِیل"
تو ہاڑا کُکڑ، تو ہاڑا سِیل، آکڑتے بآکڑ روٹیاں روٹیاں غُسّے نال پآکڑ کی تاڑیا
"اوہ ساھڑا کُکڑ"۔ ساھڑا سِیل پآکڑ فٹوفٹ بین بدلیا۔

Figure 36: Mehmood (2001), extract from page 41 (Pathwari).

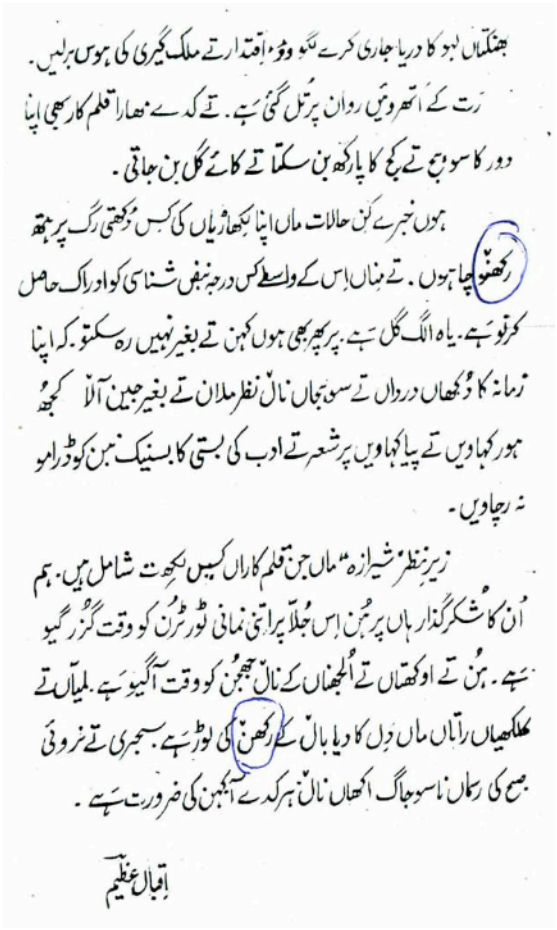


Figure 37: Azeem (1996), unnumbered page (Gojri).

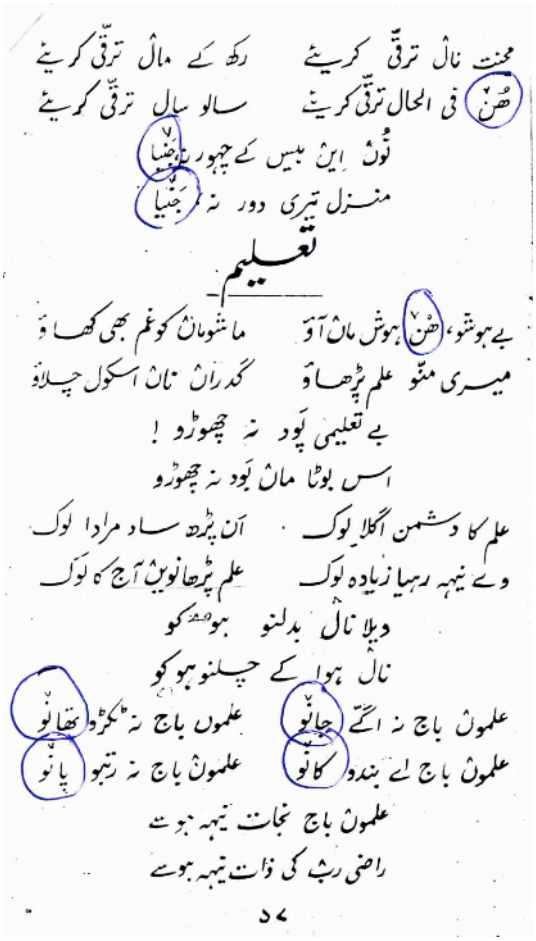


Figure 38: Afaqi (n.d.), page 57 (Gojri).

3.4 Triple-dot Arabic punctuation mark

Traditional orthographic practice when writing African languages such as Hausa, Wolof, Fulani, Mandinka, etc., in Arabic script includes the use of a characteristic punctuation mark as a "stop". This mark consists of three dots in a triangle, similar to the pattern of dots found on Arabic letters such as THEH and SHEEN, but used independently as a punctuation mark.

Although there is some suggestion that this mark is falling out of favor, with present-day writers tending to use more Latin-like punctuation, its widespread use in older texts appears to justify its encoding.

The following character is therefore proposed for addition to the UCS, with a suggested code value of U+061E:

Glyph	*Code*	*Character name*	*GC*	*CC*	*Bidi*	*See figures*
∴	061E	ARABIC TRIPLE DOT PUNCTUATION MARK	Po	0	AL	17, 39, 40, 41

The character properties proposed here are analogous to those of other Arabic-script punctuation marks such as U+061B and U+06D4.

This character should be treated similarly to the other punctuation marks in collation; it is suggested that it be given a primary weight such that it sorts after U+06D4 ARABIC FULL STOP by default, although this is a rather arbitrary choice. The collation weight can of course be tailored as needed if a different behavior is desired in any particular language.

The following figures show examples of the use of this character.

The only punctuation mark that is used is equivalent to a full stop and is composed of three dots thus - ∴
One may occasionally encounter the Arabic interrogation mark ؟ , but there is no other common form of punctuation, mainly because any punctuation as we know it can be easily confused with a letter or part of a letter in the Arabic alphabet. N.B. Writers today are tending to use the single dot full stop sign in preference to the three dot sign.

Figure 39: Addis (1963), extract from page 13.

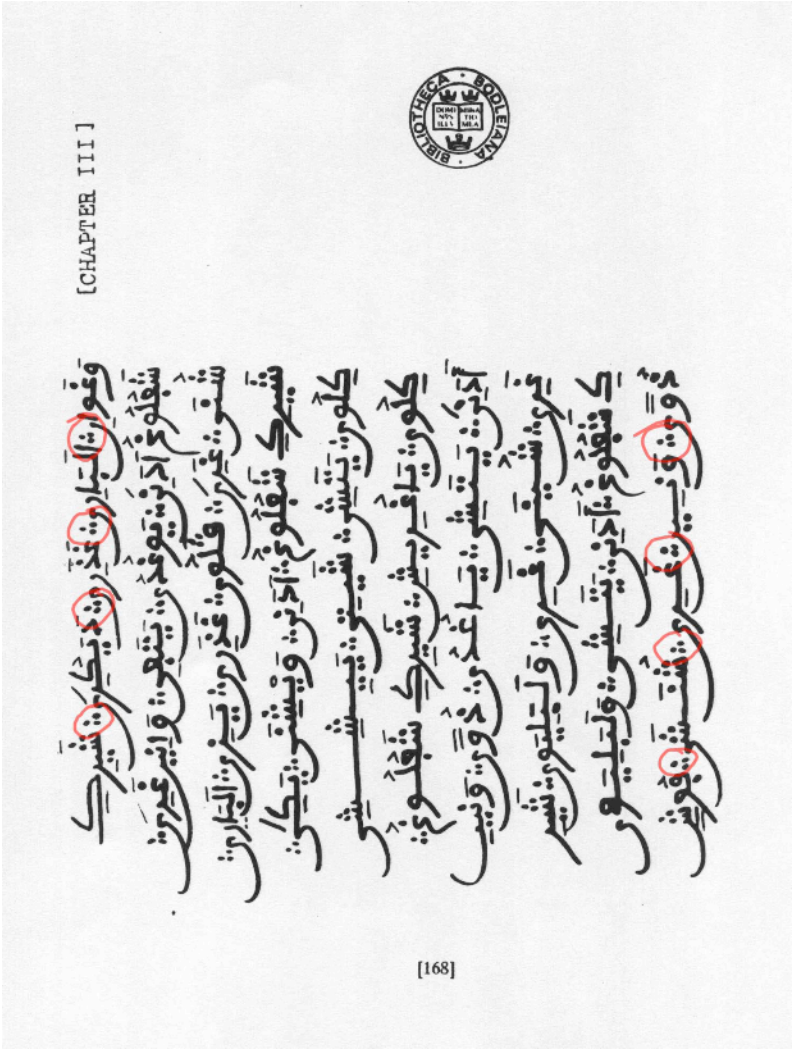


Figure 40: Piłaszewicz (1992), page 168. Many examples of the proposed character.

the syllable it is used alone in order to avoid two alifs: this is done in Arabic by the madda ~, really an alif written horizontally: ex.

ءَاىَ **āya,** the sign ▲ used as a stop

آيَة 'āyatun', is the same word in Arabic

أَلْفُرْءَانِ **Alḳura'āni,** and فُرْءَانِ **Ḳur'āni,** the Koran

Figure 41: Taylor (1929), extract from page 33.

4. References

Addis, R. T. 1963. *A study on the writing of Mandinka in Arabic script.* [SOAS library, London: call # LY Mandingo A/574203]

Afaqi, Dr. Sabir. n.d. *Pegham-in-qalab*. Pakistan. [Gojri poetry]

Akbar, Akbar Hussain. 1985. *Sumulo Rasuul (Holy Prophet)*. Islamabad, Pakistan: Modern Book Depot.

Azeem, Iqbal Chaudhry, ed. 1996. *Sheeraza*. Srinagar: Jammu & Kashmir Academy of Art, Culture and Languages. [Gojri periodical]

Bhaya, Bashir Ahmad. 1984, revised 1998. *Saraiki quaid te zubandani [Saraiki, father of languages].* Bahawalpur, Pakistan: Saraiki Adabi Majlis.

Chitka Committee. n.d. *Pathwari/Pahari nã pahala qaidah [First Pathwari/Pahari primer].* Jhelum, Pakistan: Praala Publishers.

Chtatou, Mohamed. 1992. *Using Arabic script in writing the languages of the peoples of Muslim Africa*. Rabat: Institute of African Studies. [Reports on a series of workshops held in several countries to work towards standardization of Arabic-script orthographies for major African languages.]

Dahab, Abdoulay Ali, Abdoulay Issakha, Badour Adbelkerim and Evodie Zürcher. 2002. *Kitab aafe kiraa naa [Livret sur la santé]*. Abéché: Projet de développement de la langue maba. [Simple health booklet in the Maba language of Chad, Arabic script edition.]

Daily Jhoke. Multan, Pakistan. [Saraiki daily newspaper]

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

Kew, Jonathan. 2002. *Proposal for extensions to the Arabic block.* L2/02-274.

———. 2003. *Proposal to encode Arabic triple dot punctuation mark*. L2/03-159.

———. 2003. *Proposal to encode Arabic-script letters for African languages*. L2/03-168.

———. 2003. *Proposal to encode Jawi and Moroccan Arabic GAF characters.* L2/03-176.

———. 2003. *Draft chart showing UTC #95 additions to Arabic blocks*. L2/03-210.

Mansour, Kamal. 2003. http://www.bisharat.net/A12N/Afro-Arabic_Symbols.pdf. [A collection of glyphs designed for African-language use; it is unclear in some cases whether actual use is established.]

Mehmood, Tariq. 2001. *Sayana gidhar [Clever jackal].* Jhelum, Pakistan: Praala Publishers.

Mughal, Shaukat. *Saraiki qaidah [Saraiki primer].* Multan, Pakistan: Saraiki Isha'ati Idarah.

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

Nodjindaina, Jean-Bosco, Gami Ssane Mogaye, Mbanji Bawe Ernest, Susan Rose and Matt Day (illus.). 2002. *Erniŋe kadade-naanu (Ernime) [Les animaux sauvages de la brousse]*. 2ème édition révisée. N'Djaména/Abéché: Association SIL. [Picture book of animals and birds of the Maba area, Chad.]

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

Piłaszewicz, Stanisław. 1992. *The Zabarma conquest of north-west Ghana and Upper Volta: A Hausa narrative "Histories of Samory and Babatu and others" by Mallam Abu.* Warsaw: PWN—Polish Scientific Publishers.

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

Taj, Abdul Khaliq. 1989. *Shina qaidah [Shina primer].* Gilgit, Pakistan: Muhammad Book Stall.

Taylor, F. W. 1929. *Fulani-Hausa readings in the native scripts: With transliterations and translations.* Oxford: OUP.