Proposal to encode Arabic-script letters for African languages

Date: June 2, 2003 (revised)
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 Arabic-script letters for African languages
2.	Requester's name	SIL International (contacts: Jonathan Kew, Peter Constable)
3.	Requester type	Expert contribution
4.	Submission date	May 30, 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	23
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?	No
2a.	Has contact been made to members of the user community?	Yes
2b.	With whom?	Linguists studying African languages; NGOs involved in education and community development among relevant language communities.
3.	Information on the user community for the proposed characters is included?	Yes
4.	The context of use for the proposed characters	Publications in African languages such as Fulani, Wolof, Hausa, Mandinka, Maba; scholarly and pedagogical use.
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 note in section II.1 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?	Vowel signs appear similar to generic combining marks from 03xx block.
9b.	Rationale for inclusion?	Script-specific vowel marks; inappropriate to unify across scripts. Typical appearance to be harmonized with other Arabic vowel marks, not with Latin diacritics. Combining classes and collation weights should be as other Arabic vowels.
10.	Does the proposal include the use of combining characters and/or use of composite sequences?	Yes, vowel signs are combining characters.
11.	Does the proposal contain characters with any special properties?	Base characters have right-to-left (AL) directionality; vowel signs are combining marks.

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. Background

The principal source of information concerning 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).

II. Proposal

The following characters are proposed as additions to the UCS repertoire. They are all characters that have been used in the orthographies of various African languages when written in Arabic script.

1. Base characters

The following 20 extended Arabic letters are found in the sources. All these proposed characters are of General Category Lo; Combining Class 0; Bidi Type AL. Suggested codepoints are of course subject to change.

Note that if the Arabic-script modifier marks proposed in L2/03-154 are encoded, it will then be possible to represent these letters using sequences of *base + modifiers,* and there will be no need to encode them as individual characters.

Glyph	*Code*	*Character name*	*Shaping*	*See figures*
ﻳ	0604	ARABIC LETTER BEH WITH THREE DOTS HORIZONTALLY BELOW	BEH	3, 7, 9
ﺛ	0605	ARABIC LETTER BEH WITH ONE DOT BELOW AND THREE DOTS POINTING UPWARDS ABOVE	BEH	4, 7, 9
ﺟ	0606	ARABIC LETTER BEH WITH THREE DOTS POINTING UPWARDS BELOW	BEH	4, 7, 8, 10, 11
ﺗ	0607	ARABIC LETTER BEH WITH THREE DOTS POINTING UPWARDS BELOW AND TWO DOTS HORIZONTALLY ABOVE	BEH	5
ﻧ	0608	ARABIC LETTER BEH WITH TWO DOTS HORIZONTALLY BELOW AND ONE DOT ABOVE	BEH	7, 8, 9
ﺒ	0609	ARABIC LETTER BEH WITH INVERTED SMALL V BELOW	BEH	10
ﺐ	060A	ARABIC LETTER BEH WITH SMALL V ABOVE	BEH	10
ﺣ	060B	ARABIC LETTER HAH WITH TWO DOTS HORIZONTALLY ABOVE	HAH	2, 3, 7, 8
ﭽ	0616	ARABIC LETTER HAH WITH THREE DOTS POINTING UPWARDS BELOW	HAH	8
ﮈ	0617	ARABIC LETTER DAL WITH INVERTED SMALL V BELOW	DAL	11
ﮋ	0618	ARABIC LETTER REH WITH STROKE THROUGH	REH	11
ﻋ	0619	ARABIC LETTER AIN WITH TWO DOTS HORIZONTALLY ABOVE	AIN	2, 3, 7, 9, 10, 11, 12, 13
ﻎ	061A	ARABIC LETTER AIN WITH THREE DOTS POINTING DOWNWARDS ABOVE	AIN	6, 8
ﻏ	061C	ARABIC LETTER AIN WITH TWO DOTS VERTICALLY ABOVE	AIN	8
ﻔ	061D	ARABIC LETTER FEH WITH TWO DOTS HORIZONTALLY BELOW	FEH	3, 7, 8, 9
ﭿ	061E	ARABIC LETTER FEH WITH THREE DOTS POINTING UPWARDS BELOW	FEH	11
ﮒ	063B	ARABIC LETTER KEHEH WITH THREE DOTS POINTING UPWARDS BELOW	GAF	11
ﻣ	063C	ARABIC LETTER MEEM WITH ONE DOT ABOVE	MEEM	3, 7
ﻡ	063D	ARABIC LETTER MEEM WITH ONE DOT BELOW	MEEM	12, 13
ﻥ	063E	ARABIC LETTER NOON WITH TWO DOTS HORIZONTALLY BELOW AND ONE DOT ABOVE	NOON	1, 11, 13

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 suggested here are based on the assumption that an ARABIC WASLA, as proposed in L2/03-166 (Alhonen 2003), is encoded at U+0659.

Glyph	*Code*	*Character name*	*GenCat*	*CombClass*	*BidiType*
◌ٚ	065A	ARABIC VOWEL SIGN SMALL V ABOVE	Mn	30	NSM
◌ٙ	065B	ARABIC VOWEL SIGN INVERTED SMALL V ABOVE	Mn	30	NSM
◌ٜ	065C	ARABIC VOWEL SIGN DOT BELOW	Mn	32	NSM

III. Examples

Note: These figures are also available online at *http://www.jfkew.plus.com/unicode/chtatou/*.

TABLE I

COMPARATIVE TABLE OF ARABIC AND SONGHOY CHARACTERS

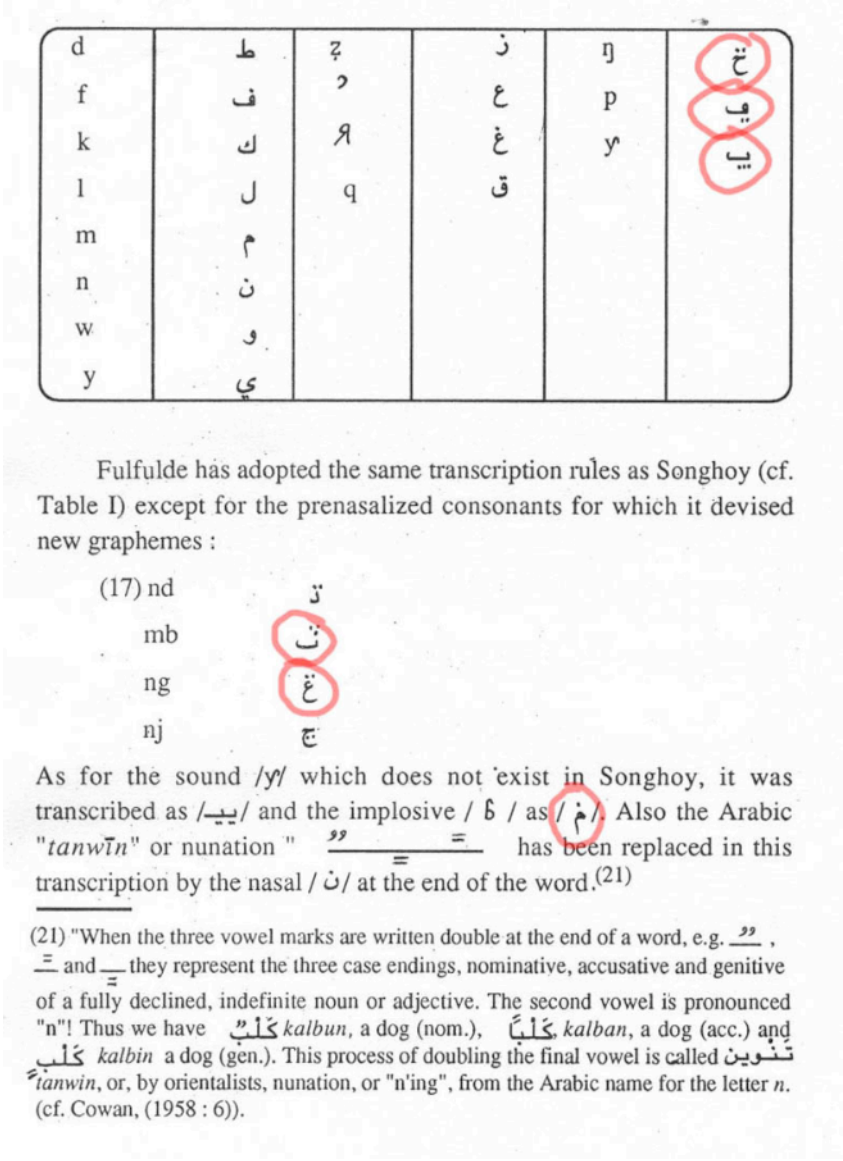
Similar characters in Arabic and Songhoy		Specific Arabic characters		Specific Songhoy characters	
Latin characters	Arabic characters	Latin characters	Arabic characters	Latin characters	Arabic characters
a	أ	θ	ث	p	ف
b	ب	ḥ	ح	ɲ	ڹ
t	ت	kh	خ	ny	ڽ

(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	ذ	ŋ	خ
d	د	ṣ	ص	g	غ
r	ر	ḍ	ض		
z	ز	ṭ	ط		
s	س	ẓ	ظ		
š	ش	ʿ	ع		
f	ف	ɣ	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	ʾ	ع	p	ڢ
k	ك	ʀ	غ	ɣ	ڢ
l	ل	q	ق		
m	م				
n	ن				
w	و				
y	ي				

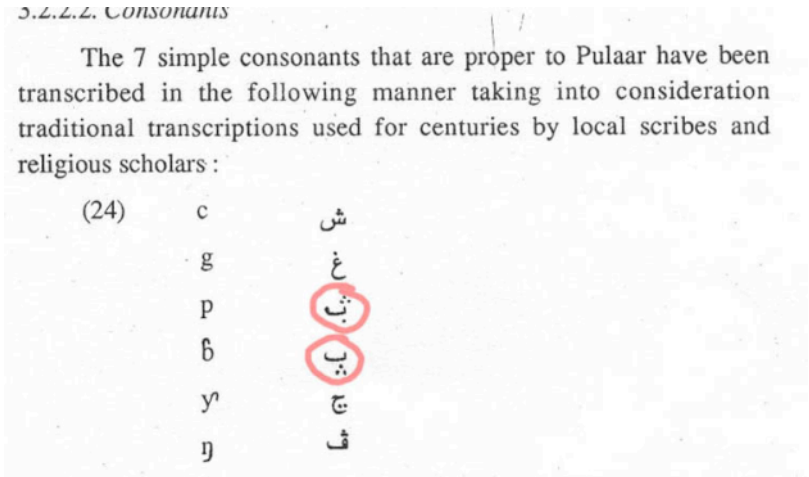
Fulfulde has adopted the same transcription rules as Songhoy (cf. Table I) except for the prenasalized consonants for which it devised new graphemes :

(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.

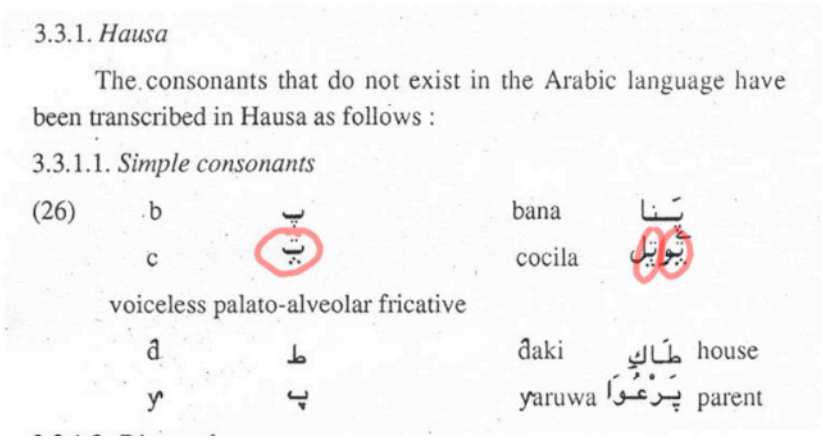


3.2.2.2. Consonants

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 assumed to be an idiosyncrasy that probably does not merit separate encoding.



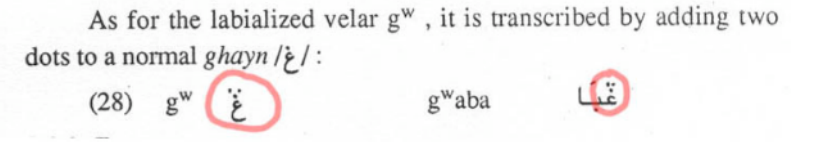
3.3.1. *Hausa*

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	ب	bana	بَـنا	
	c	چ	cocila	چُوچِيلَ	
	voiceless palato-alveolar fricative				
	ɗ	ط	ɗaki	طَـاكِ	house
	y	پ	yaruwa	پَـرْعُـوَا	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 غ g^waba غْبَا

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	ٺَكَّ - شَكَّ	necklace
g	ڠ	غ	gerte	ڠرْت غرْت	ground-nut peanut
p	ڢ	ڢ	paaka	ڢَاكَ ڢَاكَ	knife
ny	ڽ	ڽ	nyiiwa	ڽِيوَ	elephant
y	ڽ	ڽ	yiyal	ڽِييَلْ ڽِييَلْ	bone
ŋ	ڠ	ق	ŋenyema	ڠَيْمَ قيَمَ	earring
ɓ	ڢ	ڢ	ɓoolde	ڢَولْد ڢِولْد	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		Examples	Gloss	
	Mali	Niger			
c	ٺ	چ	ciiri	نَيرُ چِير	salt
g	غٺ	غ	gaara	غَارَ غَارَ	to sollicit
ny	ٺ	ٿ	nyaamoy	يَامَىْ ٿَامَىْ	
p	ڢ	ڀ	paate	ڢَات ڀَات	
ŋ	ڭ	ڠ	naari	ڭَارِ ڠَارِ	rice
o	ـٺ	ـٺ	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 ـٖ ـٗ
ت	ٿ	نج	ـ٘ ـٙ ـٚ
ج	ڠ	نع	
د	ڃ		
ر	ڼ		
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)

ٽ for **c** (instead of ث)

ج for **ɣ** (no change)

ط for **ɗ** (no change)

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

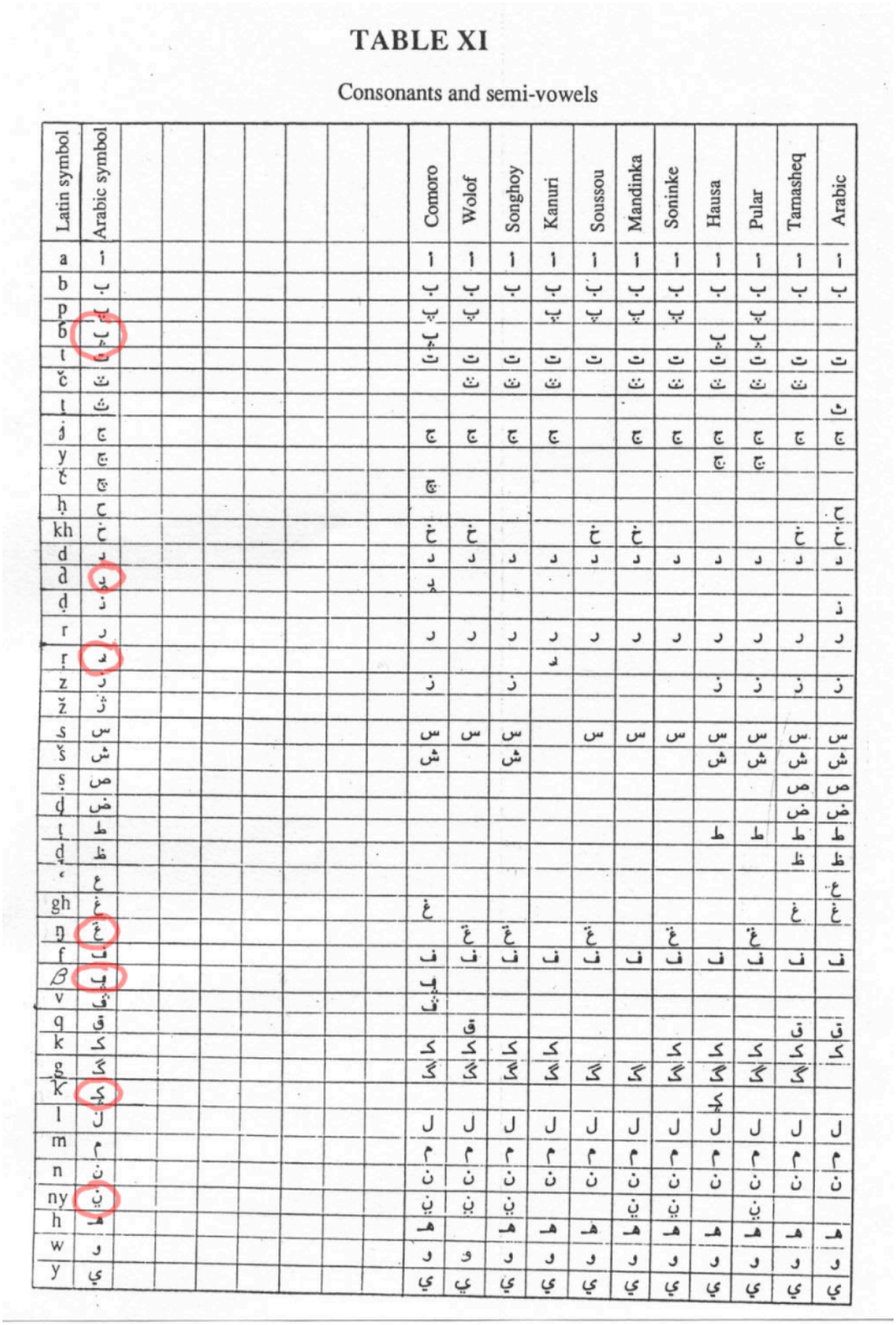
ڠ for **ŋ** (instead of ڤ)

ـٖ for the vowel **e**

Figure 10: Chtatou (1992), page 54.

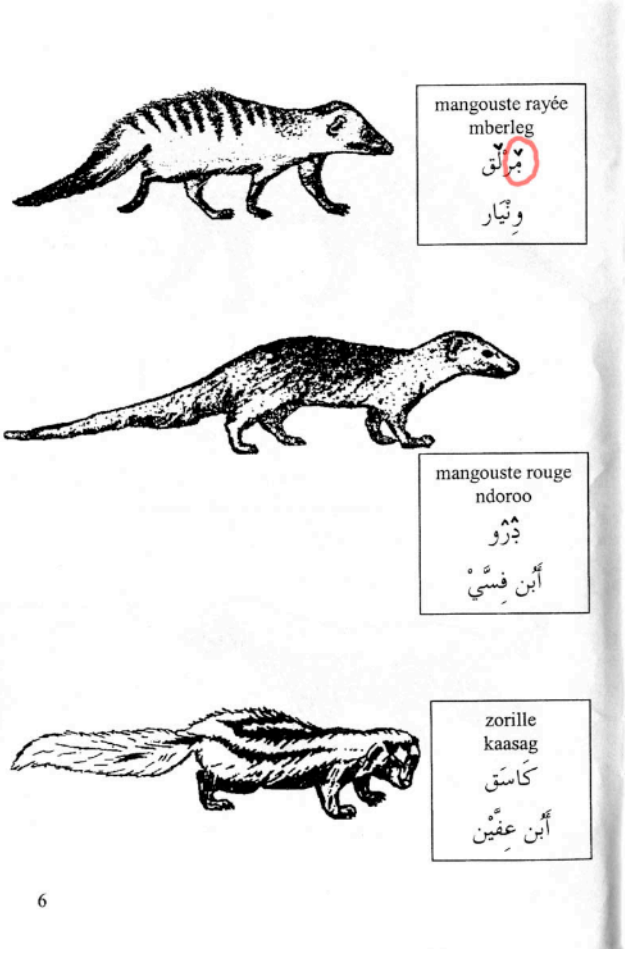
TABLE XI

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		
ɓ	ɓ	ɓ							ɓ	ɓ		
t	ت	ت	ت	ت	ت	ت	ت	ت	ت	ت	ت	ت
č	ث		ث	ث	ث		ث	ث	ث	ث	ث	
ṯ	ث											ث
j	ج	ج	ج	ج	ج		ج	ج	ج	ج	ج	ج
y	ج								ɲ	ɲ		
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
ɗ	ɗ	ɗ										
ḍ	ض											ض
r	ر	ر	ر	ر	ر	ر	ر	ر	ر	ر	ر	ر
ṛ	ɽ				ɽ							
z	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
ḍ	ض										ض	ض
ṭ	ط								ط	ط	ط	ط
ḍ	ظ										ظ	ظ
ʿ	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
ɸ	ڤ	ڤ										
v	ڤ	ڤ										
q	ق		ق								ق	ق
k	ك	ك	ك	ك	ك			ك	ك	ك	ك	ك
g	گ	گ	گ	گ	گ	گ	گ	گ	گ	گ	گ	
ǩ	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	ڼ	ڼ	ڼ	ڼ			ڼ	ڼ		ڼ		
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.



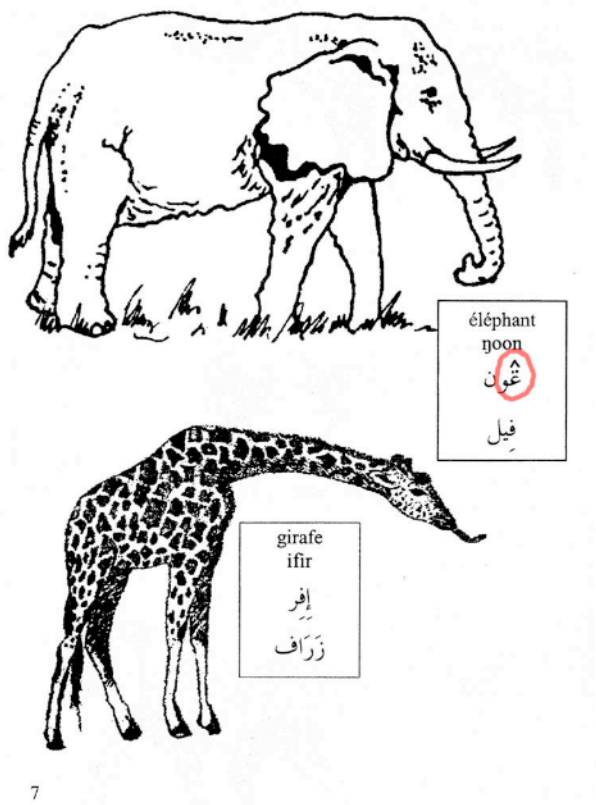
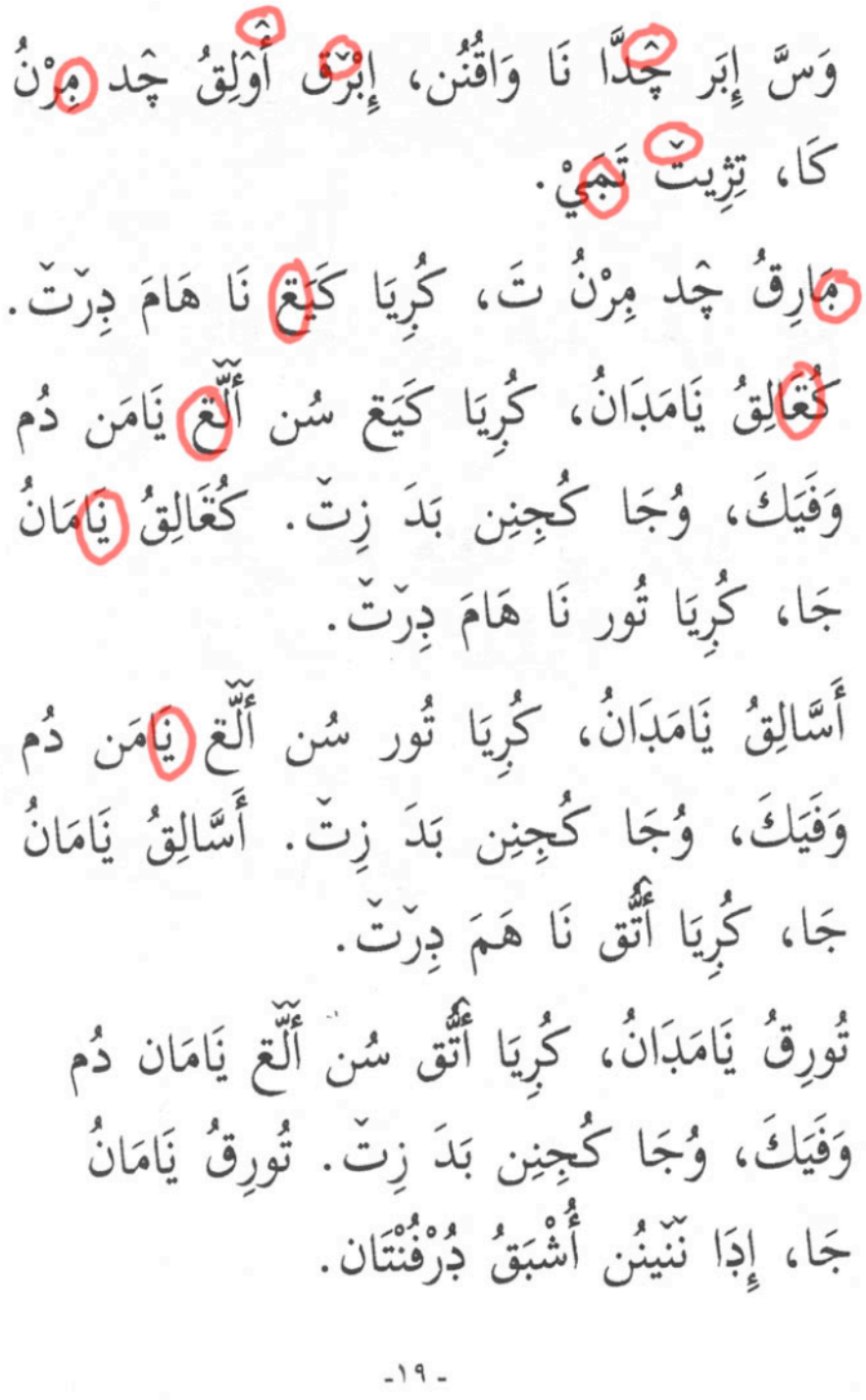


Figure 12: Nodjindaina et al (2002), pages 6–7.


وَسَّ إِبَر چْدَّا نَا وَاقُنُن، إِبْرَق أَوْلِقُ چْد مِرْنُ
كَا، تِزِيتّ تَمِيْ.
مَارِقُ چْد مِرْنُ تَ، كُرِيَا كَيَع نَا هَامَ دِرّتّ.
كُغَالِقُ يَامَدَانُ، كُرِيَا كَيَع سُن أَلّغ يَامَن دُم
وَفَيَكَ، وُجَا كُجِنِن بَدَ زِتّ. كُغَالِقُ يَامَانُ
جَا، كُرِيَا تُور نَا هَامَ دِرّتّ.
أَسَّالِقُ يَامَدَانُ، كُرِيَا تُور سُن أَلّغ يَامَن دُم
وَفَيَكَ، وُجَا كُجِنِن بَدَ زِتّ. أَسَّالِقُ يَامَانُ
جَا، كُرِيَا أَتّق نَا هَمَ دِرّتّ.
تُورِقُ يَامَدَانُ، كُرِيَا أَتّق سُن أَلّغ يَامَان دُم
وَفَيَكَ، وُجَا كُجِنِن بَدَ زِتّ. تُورِقُ يَامَانُ
جَا، إِدَا نّنِنُن أُشْبَقُ دُرْفُنْتَان.

-١٩-

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

IV. Draft chart showing proposed additions to Arabic block

V. References

Alhonen, Miikka-Markus. 2003. *Proposal for encoding the combining diacritic ARABIC WASLA*. L2/03-166. [See draft code chart, section IV.]

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.]

Kew, Jonathan. 2003. *Proposal to encode Arabic triple dot punctuation mark*. L2/03-159. [See draft code chart, section IV.]

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

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.]

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.]