

Some comments on the Arabic block in Unicode

Tom Milo, DecoType

Summary

1. Some Extended Arabic characters are typographical variants of characters already adequately covered by the corresponding Basic Arabic Characters;
2. 0626 ARABIC LETTER YEH WITH HAMZA ABOVE is actually a representation form of two nominal characters: 06D5 ARABIC LETTER AE followed by 0621 ARABIC LETTER HAMZA;
3. The graphemes for the aspirated phonemes of Urdu should be added; 06BE ARABIC LETTER HEH DOACHASHMEE can then be deleted or ignored;
4. Urdu Noon-e-ghunna needs fourfold display forms;
5. The characteristic Urdu digits are missing;
6. The Ottoman *Kaf-i-Turki* (with additional stroke below the main tail) is missing
7. 0644 0644 0647 ARABIC LIGATURE LI-LLAH ISOLATED FORM is missing

1. Doublets: *Teh Marbuta, Feh, Qaf, Kaf, Heh, Yeh*

In some cases it appears to me that Unicode - knowingly - confuses regional calligraphic or typographic variants for encodable characters. E.g., the encoded HEH GOAL, along with its associated GOAL variants, is clearly an attempt to tweak the *Naskh* typeface (as used for printing the Unicode Standard) to look like *Nastaliq*: an Eastern variant of the Arabic script.

Characters concerned:

Basic Arabic	Extended Arabic doublets
0629 ARABIC LETTER TEH MARBUTA	06C3 ARABIC LETTER TEH MARBUTA GOAL
0641 ARABIC LETTER FEH	06A2 ARABIC LETTER FEH WITH DOT MOVED BELOW
0642 ARABIC LETTER QAF	06A7 ARABIC LETTER QAF WITH DOT ABOVE
0643 ARABIC LETTER KAF	06A9 ARABIC LETTER KEHEH
0647 ARABIC LETTER HEH	06C1 ARABIC LETTER HEH GOAL
064A ARABIC LETTER YEH	06CC ARABIC LETTER FARSI YEH

In fact, this "goal" effect is not at all obligatory for Naskh, as is illustrated here (second row: *he-yi-hewwez*):

Название букв	Транскрипция	Формы букв			
		изолированно	в конце соединения	в середине соединения	в начале соединения
«хэ»	h	ح	ح	ح	ح
«ха-йе-хавваз»	h	ه	ه	*هـ — هـ	هـ — هـ
«гайн»	g	غ	غ	غ	غ
«кәф»	k	ق	ق	ق	ق

Illustration taken from S. K. Gorodnikova and L.B. Kibirskhtis, *Uchebnik Jazyka Urdu (Urdu Primer)*, Moscow 1969., Moscow 1969. Also compare the names of numbers 6 and 8 as written in Nastaliq or printed in Naskh in the illustration accompanying the paragraph on Urdu-Indic numbers.

The asterisk following the second middle heh leads to a note pointing out that the alternative form of the letter is used to create aspirated consonants¹.

The same is true for KAF and so-called KEHEH (06A9): they represent one and the same Arabic letter KAF (0643). As for the name *keheh*: as far as I know it is in use Urdu to denote the aspirated phoneme /k^h/ which still lacks a proper character code in Unicode.

I believe that the correct treatment of Urdu HEH and KEHEH is to use the regular Arabic HEH and KAF with a properly designed font, in casu Nastaliq, to render the "user-expected" shape.

I observed the same phenomenon in 06CC FARSI YEH to complement 064A ARABIC YEH. Arabic YEH with or without dots in final position is a matter of regional and stylistic preference: e.g., in traditional Egyptian typesetting and calligraphy YEH in final and isolated position never had dots. In Persian this traditional style is still the only one allowed, therefore final and isolated YEH with dots do not occur.

Use of Magribi variants of FEH and QAF rules out the use of the corresponding Middle Eastern variants of these letters in the same context. They are not entitled to Unicodes and should be dealt with by font designers.

I believe using regional flavours of fonts is acceptable, since the differences in are well known to the users and do not bar him or her from understanding raw text.

2. Heh with Yeh

Heh with Yeh in Unicode represents the Arabic *letter sequence* that is associated with a syntactic construction called *izafe* or *izafet*, the Persian equivalent of nominal word composition or linking. It occurs also in Tajiki, Pashtu, Dari, Urdu and Ottoman Turkish.

A famous example is the name of the British royal diamond: *kuh-i-nur* کوه نور, literally *mountain-of-light*. The *-i-* in the example is the connecting element, paraphrased here as *-of-*. In Arabic script such a linking is optionally expressed by 0650 ARABIC KASRA. The "yeh-above" described in

Unicode 06C0 ARABIC LETTER HEH WITH YEH ABOVE corresponds to the Persian *hemze-yi-muleyyine*: the "relaxed hamza"². It is the ARABIC LETTER HAMZA used when a word ending in the vowel /e/ (best written with 06D5 ARABIC LETTER AE) is connected to the following word.

In contemporary Persian the old pronunciation of *hiatus* or glottal stop (in Arabic: *hamz*) between certain vowels has evolved into a glide /y/³. For the same phenomenon Ottoman grammarians use the Persian term *hemze-i-izafet* (still with *-i-* instead of *-yi-*): the "hamza-of-linking"⁴.

Any noun or adjective can be the first element of *izafe*. Combining AE and HAMZA in one code is just as erroneous as was the combined LAM-ALEF code point of early Arabic code pages. Furthermore, older and conservative modern spelling use the *hemze-yi-muleyyine* also on top of YEH. From this it follows that the proper solution would be to consider this floating hamza a separate character, so that users have the freedom to use modern or historical spellings. Needless to say that it would also help Ottoman Turkish data processing.

To control the positioning in fonts, designers can still substitute a ligature.

All of this also applies to 06C2 ARABIC LETTER HEH GOAL WITH HAMZA above (which happens to be pronounced exactly like modern Persian: /yi/).

3. Aspirated consonants in Urdu and Hindi

The real *Keheh*, whose name is, mistakenly, used in the description of the Persian and Urdu doublets of Kaf, is in fact member of a class of aspirated phonemes that with the present encoding can only be encoded by combining the letter with 06BE ARABIC LETTER HEH DOACHASMEE.

I propose to replace this composition method with the proper graphemic encodings, following Hindi practice.

Urdu and Hindi are closely related languages, if not one and the same language spoken in different cultures, i.e., Islam and Hinduism. Their phonological systems share the distinctive feature of *aspiratedness*. An authoritative publication like *The Worlds Major Languages*⁵ deals with both languages in one chapter: Hindi-Urdu. It gives the following consonant scheme (*asp.* stands for aspirated):

Consonants			<i>Labial</i>	<i>Dental</i>	<i>Retroflex</i>	<i>Alveo-Palatal</i>	<i>Velar</i>	<i>Back Velar</i>
Stop	vls.	unasp.	p	t	ʈ	č	k	(q)
		asp.	ph	th	ṭh	čh	kh	
Nasal	vd.	unasp.	b	d	ɖ	ǰ	g	
		asp.	bh	dh	ɖh	ǰh	gh	
Flap	vd.	unasp.			ɾ	r		
Lateral		asp.			ɽ			
Fricative	vls.		(f)	s	(ʃ)	š	(x)	
	vd.			(z)		(ž)	(ɣ)	
Semi-vowels			w (v)			y		

Both Hindi and Urdu can be traced to Sanskrit, the classic language of Hinduism. The discovery in 18th century of this language and its literature lead to foundation of modern linguistic thinking in Europe. On the one hand the realization of its similarity to other classic languages like Greek and Latin signalled the beginning of Comparative Linguistics and General Linguistics with all its consequential spin-offs like the Neo-grammarians, Structuralism, the Prague School, even Generative Transformationalism and, the latest in comparative linguistics, the Nostratic Theory.

On the other hand, Sanskrit was not just a passive object of study, it also actively contributed the discipline of phonetics, a key factor in the emergence of modern linguistic thinking⁶. Against this backdrop it should come as no surprise that Devanagari script is the most accurate phonetic writing system known in history. The Hindi writing system with Nagari inherits from Sanskrit a very precise orthography with a virtual one-to-one relation between phonemes and graphemes. Consequently it recognizes the aspirated consonant phonemes as independent graphemes⁷:

	<i>Voiceless unaspirated plosives.¹</i>	<i>Voiceless aspirated plosives.¹</i>	<i>Voiced unaspirated plosives.¹</i>	<i>Voiced aspirated plosives.¹</i>	<i>Nasals.</i>
Velars	क <i>ka</i>	ख <i>kha</i>	ग <i>ga</i>	घ <i>gha</i>	ङ <i>ṅa</i>
Pre-palatals	च <i>ca</i>	छ <i>cha</i>	ज <i>ja</i>	झ <i>ḥja</i>	ञ <i>ña</i>
Retroflexes	ट <i>ṭa</i>	ठ <i>ṭha</i>	ड <i>ḍa</i>	ढ <i>ḍha</i>	ण <i>ṇa</i>
Dentals	त <i>ta</i>	थ <i>tha</i>	द <i>da</i>	ध <i>dha</i>	न <i>na</i>
Labials	प <i>pa</i>	फ <i>pha</i>	ब <i>ba</i>	भ <i>bha</i>	म <i>ma</i>
Semivowels, etc.	य <i>ya</i>	र <i>ra</i>	ल, ल ² <i>la</i>	व <i>va</i>	
Sibilants		श <i>śa</i>	ष <i>ṣa</i>	स <i>sa</i>	
Glottal			ह <i>ha</i>		
Flaps		ड़ <i>ṛa³</i>	ढ़ <i>ṛha³</i>		

Various publications treat these aspirated phonemes as independent graphemes also in Urdu. First there is a comprehensive study on Arabic writing systems that notices the same one-to-one relationship between phonemes and graphemes as in Hindi⁸:

2.2.3.1.1. Das Phonem- und Graphemsystem des Urdu

a. Konsonantenphoneme

p	p ^h	t	t ^h	ṭ	ṭ ^h	ʃ	ʃ ^h	k	k ^h	q
b	b ^h	d	d ^h	ḍ	ḍ ^h	ʒ	ʒ ^h	g	g ^h	
m	m ^h	n	n ^h							
		l	l ^h							
		r	r ^h	ɽ	ɽ ^h					
f		s				ʂ		x		
v		z				ʐ		ɣ		ɦ
						j				

The same book continues with this grapheme table:

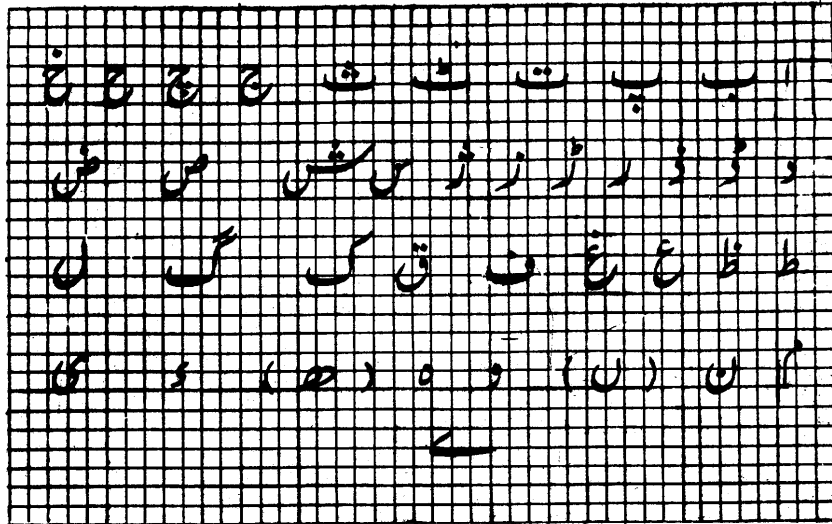
b. Konsonantengrapheme

پ	پھ	ت(1)	ٹھ	ٹھ	چ	چھ	ك	كھ	ق
ب	بھ	د	دھ	ڈھ	ج	جھ	گ	گھ	
	م	ن	نھ						
		ل	لھ						
		ر	رھ	ڑ	ڑھ				
ف		س(2)			ش		خ		
و		ز(3)			ژ		غ		ہ(4)
					ی				

The aspirated consonants are composed by writing the letter of the plain consonant followed by the *two-eyed heh*⁹. What makes this table interesting is, that, unlike regular grammars that deal with the script only cursorily and leaving a lot of questions, this book shows how well-adapted the Arabic writing system is for Urdu.

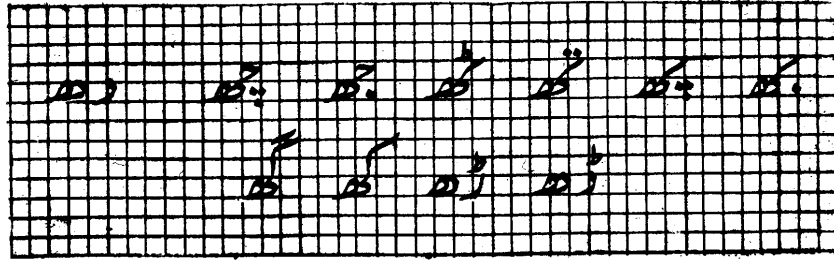
But the ultimate argument to treat aspirated Urdu consonants as independent graphemes comes from an authoritative Urdu scholar. Professor Mohammed Zakir, in his *Lessons in Urdu Script*, also begins with giving the traditional, essentially Arabic-Persian alphabet table:

THE URDŪ ALPHABET



Then he adds the following table recognizing , like Mohammad-Reza Majidi, the graphemic status of the combined letters¹⁰:

LETTERS DEVISED FOR ASPIRATED CONSONANTAL SOUNDS IN URDU



This supports the observation I am making: there should be a series of aspirated consonantal graphemes added to the Basic Arabic block in Unicode. The safest model to follow is the set given by Majidi, as it includes combinations that may be not graphemes representing phonemes, but that nevertheless seem to be recognized as such.

From this it also follows, that 06BE ARABIC LETTER HEH DOACHASHMEE can be deleted from Unicode or at least ignored.

As a result a much cleaner subset for Urdu can be created without ambiguities that such as between 0647 ARABIC LETTER HEH, 06BE ARABIC LETTER HEH DOACHASHMEE (misplaced representation form) and 06C1 ARABIC LETTER HEH GOAL (doublet), since the latter two can both be deleted or ignored.

In order to satisfy user expectation, ARABIC LETTER HEH DOACHASHMEE can still feature on the keyboard layout to allow the user to construct the real Unicodes in a way that may even turn out to be intuitive.

Finally. I believe that 06C2 ARABIC LETTER HEH GOAL WITH HAMZA ABOVE is totally redundant, as HAMZA ABOVE is in fact the floating hamza of *izafe* (para 1), and HEH GOAL is a doublet.

For Microsoft, with its already over-crowded Windows Code Page 1256 ARABIC LETTER HEH DOACHASHMEE could be maintained. It would mean that, for the sake of economy, this code page keeps track of the keyed-in sequences, rather than storing the proper grapheme codes. When converting to Unicode the resulting sequences of [plain letter] plus [aspiration mark] must, of course, be replaced by the proper codes.

4. Noon Ghunna also has non-final forms

06BA ARABIC LETTER NOON GHUNNA is given two representation forms:

FB9E ARABIC LETTER NOON GHUNNA ISOLATED FORM

FB9F ARABIC LETTER NOON GHUNNA FINAL FORM

However, it can be documented to have also an initial and a middle form with an optional distinction mark from the regular NOON in these positions. The first illustration shows the four positional variants, the initial and middle of which are identical to those of the regular NOON:

ПИСЬМЕННОЕ ВЫРАЖЕНИЕ НОСОВЫХ ГЛАСНЫХ ЗВУКОВ

Все носовые гласные изображаются на письме буквой ن «ну-гунна», которая ставится после знака соответствующего чистого краткого, или долгого гласного, или дифтонга. Эта буква имеет четыре начертания, но с нее слова никогда не начинаются. В начале или в середине соединения над ней ставится точка, как и над буквой ن «нуль».

Таблица 8

Название буквы	Транскрипция	Формы буквы			
		изолированно	в конце соединения	в середине соединения	в начале соединения
«ну-гунна»	~	ن	ن	ن	ن

Illustration taken from S. K. Gorodnikova and L.B. Kibirskhtis, *Uchebnik Jazyka Urdu*, Moscow 1969. Translation:

Nasal vowels in writing. All nasal vowels are rendered in writing using the letter *noon ghunna*, which is placed following the character representing the plain short or long vowel or diphthong. This letter (*noon ghunna* TM) has four written variants, although it can never occur in word-initial position. When connected in initial or middle position a dot is placed above it, as is the case with the letter *noon*.

The second illustration shows the same set with, in non-final positions, an additional distinctive element reminding of a breve mark:

Para 2. When nasalisation is needed medially in a word the نُقْطَة
(*nuqṭa* = dot) of ن is placed but over it is introduced the mark of نُونٌ غُنَّةٌ (ن) and it continues to be pronounced as the letter *n* in the English words *sink* and *song*. The mark of *nūn-e-ghunna*, it will be noticed, is like a semi-circle opening upwards. Thus بَنْگَال (anḡrez = the English people), بَنْگَال (*baṅgāl* = Bengal), بَانَسْرِي (*bānsrī* = the flute), پانچ (*pāñch* = five), جَنْگَل (*jaṅgal* = a wood; forest), چاند (*chānd* = the moon), کَنْوَل (*kañval* = lotus), لَنْكَا (*lañkā* = Ceylon). مِينْدَاک (*meñḡak* = a frog), سَانپ (*sāñp* = a snake) etc.

Sample taken from Mohammed Zakir, *Lessons in Urdu Script*, Delhi 1973. Also compare the names of number 5 as written in Nastaliq or printed in Naskh in the illustration accompanying the paragraph on Urdu-Indic numbers.

5. Missing? The Urdu-Indic numbers

The series of EXTENDED ARABIC-URDU DIGITS differs from Persian and Arabic. Yet it is nowhere mentioned in the Standard.

Names		Figures		1	ایک	ек	۱
<i>ek</i>	ایک	1	۱	2	دو	до	۲
<i>tin</i>	تین	3	۳	3	تین	тйн	۳
<i>chār</i>	چار	4	۴	4	چار	чār	۴
<i>pāñch</i>	پانچ	5	۵	5	پانچ	пāч	۵
<i>chhāe</i>	چھ	6	۶	6	چھ	чhā	۶
<i>sāt</i>	سات	7	۷	7	سات	sāt	۷
<i>āṭh</i>	آٹھ	8	۸	8	آٹھ	āṭh	۸
<i>nau</i>	نو	9	۹	9	نو	нао	۹
<i>daś</i>	دس	10	۱۰	10	دس	дас	۱۰
sample taken from Mohammed Zakir, <i>Lessons in Urdu Script</i> , Delhi 1973.				sample taken from S. K. Gorodnikova and L.B. Kibirskhtis, <i>Uchebnik Jazyka Urdu</i> .			

The digits *four*, *six*, *seven* and *nine* are markedly different and therefore justify a separate range just like the Persian (Arabic-Indic) digits.

On the other hand, I would prefer an approach where all digits should be merged into one series, using regional flavours of fonts instead, since differences in shape of the digits (Arabic, Persian and Urdu) do not bar a user from understanding raw text.

6. The Ottoman Kaf-i-Turki is missing

In the late Ottoman era attempts were made to simplify the spelling of Turkish with Arabic letters. From this period stem the Kaf variants with added distinctive features, elaborating the exact graphemic function of the letter concerned. Two of them happen to be covered already in the Unicode, but one seems to have been overlooked, the so-called Turkish kaf, since it seems only to be used there in etymological spellings of the phoneme /y/. Handling Ottoman archives is a serious concern for researchers and governmental institutions alike. Here are some illustrations taken manuals dealing with Ottoman Turkish writing:

آگه . آگری . دگیل . دگیمن . دکنک ایگه . ایگده . اوگله .
اوگرنک . اوگرتک . بک . بگنک . آگنجه . یگری . لکک . بورهک .
بورهکی . بورهکه . اریک . اریگی . اریکه . بشیک . بشیکی . بشیکه . اینک . اینگی .
اینکه . ده کیل بک . دوکون یگری . لکک لریووا یاپدیله .
یاورولری بیسه یورلر . بوده کیمن روزگار ایله دونه ر . ده کیمنده .
بوغدا یی اوگودورلر . اون یاپارلر . فورونجی . اوندن ائک یاپار .

Okunuşu: 1 eye, eyri, deyil, deyirmen, deynek, iyne, iyde, öyle, 2 öyrenmek, öyütmek, bey, beyenmek, eyvence, yirmi, leylek, börek 3 böreği, böreğe, erik eriği eriğe, beşik beşiği beşiğe, inek ineği 4 ineğe, değil, bey, düğün, yirmi, leylekler yuva yaptılar 5 yavrularını besleyorlar. Bu değirmen rüzgar ile döner. Değirmende 6 buğdayı öğüdürler. Un yaparlar Fırını undan ekmek yapar.

(from Nahit Tendar & Nebahat Karaorman, *Osmanlica okuma anahtari* [key for reading Ottoman], Istanbul 1970)

7. Arabic ligature li-llah

The glyph ARABIC LIGATURE LI-LLAH ISOLATED FORM and ARABIC LIGATURE LI-LLAH FINAL FORM [both representing 0644 0644 0647] are missing from the block of Arabic Representation Forms A.

This is a major defect: apart from LAM-ALEF لا, LI-LLAH لله is arguably the only *real* ligature of the Arabic script¹¹, since the writing of God's name in Naskh and Naskh-related scripts requires a LAM of reduced height.

All Microsoft fonts - with the exception DecoType supplied fonts, that use the Private Area prescribed by the Unicode Standard - apparently assume there must have been some mistake and correct this error partly by replacing the redundant FDF2 [0627 0644 0644 0647] ARABIC LIGATURE ALLAH ISOLATED FORM with FDF2 [0644 0644 0647] ARABIC LIGATURE LI-LLAH ISOLATED FORM. This replacement is justifiable, since the word ALLAH الله consists of the ligature LI-LLAH لله preceded by a standard ALEF ا.

The Microsoft workaround may turn out to be a rather useful (de facto) standard as it leads to the desired effect in nearly all contexts¹².

Notes

¹ Though this is true, this same alternative form can still be used for writing the independent phoneme /h/. Cf. Mohammed Zakir:

The aspirated consonantal sounds have no separate letters in Urdû; these are obtained by using **ھ** *do chashmi* (two-eyed) **ه** at the end of certain consonants. Thus **ھ** *do chashmi* **ه** may be called the aspirated consonant-forming device. (It should be noted, however, that sometimes it is employed as a medial form of the letter **پ** *he* as introduced in Lesson 7. The ordinals **بارھواں** , **بارھوپ** (*bārhvān*, *bārhvīn* = twelfth) and **تیرھواں** , **تیرھوپ** (*terhvān*, *terhvīn* = thirteenth) derived from words ending on **پ** *he* viz. **بارہ** (*bārah* = twelve) and **تیرہ** (*terah* = thirteen) respectively illustrate this point. These ordinals should have been written as **بارہوپ** , **بارہواں** and **تیرہوپ** , **تیرہواں** respectively employing the form **پ** of the letter **پ** , as introduced in Lesson 7 Para 2, but there is a general practice to write these with **ھ** *do chashmi* **ه***)

² First mentioned in a 12-13th century Persian work **المعجم فی معایر الاشعار الہج** but also quoted in recent publications like Prof. M. Moin's *Izafa - the genitive case*, Teheran 1984 (information communicated to me by Dr H.U. Qureshi, former head of the of the Persian Department of the Jamia Millia Islamia, New Delhi, later lecturer of Persian at the University of Tehran and for many years coordinator of the Language Services for the Iran-United States Claims Tribunal in The Hague)

³ cf. Gilbert Lazard, *A grammar of Contemporary Persian*, New York 1992, pp. 32-33:

§22. It frequently happens that two vowels are found next to each other at the junction of a word and of an enclitic. The treatment depends in this case not only on the quality of the vowels which are side by side, but also on the nature of the enclitic. It often varies, besides, according to the stylistic register.

2) "Ezâfe" *e-* (§44): except after *i*, *-e* is regularly preceded by a *y* of transition:

Ex. *zendegi-e Hâfez* "the life of Hâfez," *dârû-ye sarvat* "endowed with wealth," *xâne-ye (xune-ye) Hasan* "the house of Hasan," *bu-ye gol* "the perfume of the flower," *har do-ye šomâ* "you two" (lit., both of you).

(note that the transcription uses the letters *i* and *a* for the long vowels /i/ and /a/ respectively; and the letter *e* to represent two different short vowels: /e/ and /i/)

⁴ This term is used by Prof. Dr Faruk Timurtash (sorry, s-cedilla not supported!) in his *Ottoman Turkish Grammar* (Istanbul 1985), in the chapter on Persian elements in Ottoman Turkish. Please note that in Ottoman Turkish the older Persian practice survives: the glottal stop // of the hamza is not yet replaced by the modern Persian glide /y/. Cf. page 260:

V			IV	III	II	I
Exemplifications			Trans-literation Or Power	Names	Aspirated Sounds and their composition	Serial No.
Final	Medial	Initial				
شبه	سبھا	بھارت	bh	bhae	بھ (ب + ہ)	1
-	پبھا	پھل	ph	phae	پھ (پ + ہ)	2
نتھ	کتھا	تھال	th	thae	تھ (ت + ہ)	3
پٹھ	بیتھک	ٹھا کر	th	thae	ٹھ (ٹ + ہ)	4
سبجھ	بجھانا	جھیل	jh	jhae	جھ (ج + ہ)	5
رچھ	بچھڑا	چھلنی	chh	chhae	چھ (چ + ہ)	6
بڈھ	بڈھن	ڈھن	dh	dhae	ڈھ (د + ہ)	7
گڈھ	مڈھا	ڈھال	ḡh	ḡhae	ڈھ (ڈ + ہ)	8
اساڑھ	گڑھا	-	rḡh	rḡhae	ڑھ (ڑ + ہ)	9
اپکھ	پتکھا	کھیل	kh	khae	کھ (ک + ہ)	10
سنگھ	سنگھی	گھری	gh	ghae	گھ (گ + ہ)	11

¹¹ All the other so-called ligatures can be considered regular calligraphic mergers of letter groups.

¹² However, final forms of li-llah are usually ignored: e.g., MS Traditional Arabic uses the same skeleton for both fa-li-llah "and to/for God" and qallallahu "he reduced it":
قللہ - قللہ

As a comparison the output of DecoType ACE (Arabic Calligraphic Engine): قللہ - قللہ