

To: UTC
Title: Regarding the Sindhi Heh
From: Lorna Priest Evans (SIL International)
Date: 6 July 2021

Introduction

The Sindhi language has two aspirated consonants (/dʒ^h/ and /g^h/) and a normal /h/ consonant. Over the years, the design of the normal /h/ consonant (in Arabic script) has been a source of confusion for many. Neither the HEH nor the HEH DOACHASHMEE provide the exact glyph shapes desired by many Sindhi language users. Additionally, there is a /ə/, /ə^h/, or no sound at all in word final position (typically represented by U+06C1 HEH GOAL). There are also a number of borrowed Arabic words where the typical shapes for HEH are expected.

Currently SIL International and the Sindhi Language Authority have two separate solutions that make any data files incompatible with the other solution. This not only has an impact on fonts, but it also impacts keyboard output. Thus, it is important to have a solution so that all documents can begin following the same standard. The aim of this document is to come to consensus on which codepoints and glyph design to use for the Sindhi *heh* and aspirated consonants. It should also be noted that although Sindhi is the most well known language using these glyphs, there are other languages requiring them as well: Kachi [gjk], Dhatki [mki], Marwari [mve], Oadki [odk], Thayadari /Wadiyari [kxp], Saraiki [skr], and Parkari [kvx].

The following table illustrates what characters and glyphs (shapes, not codepoints) are required for Sindhi. A standard Sindhi /h/, an Arabic /h/, /dʒ^h/, /g^h/, and /ə/ or /ə^h/ all seem to be required (when Sindhi is written in Devanagari script, it has 3 different codepoints for the characters being discussed in Arabic script).

Table 1: Sindhi character requirements

	Transliteration	fina	medi	init	isol		Sindhi in Devanagari script
1.1	/h/ -1	ھ	ھ	ھ	ھ		ह (U+0939 HA)
1.2	/h/ -2	ہ	ہ	ہ	ہ	Used for borrowed words	
1.3	/ɟʰ/	جھ	جھ	جھ	جھ		झ (U+091D JHA)
1.4	/gʰ/	گھ	گھ	گھ	گھ		घ (U+0918 GHA)
1.5	/h/ (other aspiration)	ھ	ھ	-	-		
1.6	/ə/, /əʰ/, short vowel, or no sound at all	ـ	-	-	ہ		

Table 2: Examples of Sindhi words using these characters

		fina	medi	init	isol	
2.1	/h/ -1	ويھُ	مهينن	هو	دوھ	
		wehʉ (sit down)	məhinən (months)	ho (was)	ḡohe (both)	
2.2	/h/ -2	وآله			وَحْدَهُ لَا	Borrowed Arabic words
		və-alah (I swear God)			wəhʉḡəh la (one alone /no one)	
2.3	/ɟʰ/	ڪُجھ	منجهان	اُجھي	باجھ	
		kudʒʰ (a few)	məndʒʰā (of them)	ʉɟʰi ʉɟʰi (meet secretly)	baɟʰə (mercy)	
2.4	/gʰ/	سگھ	گھڻگھرن	گھوٽ	گھ	
		səgʰə (strength)	gʰəŋ-gʰʉrən (well wisher)	gʰoʈʉ (groom)	gʰe (wheat)	
2.5	/h/ (other aspiration)	ٿاھ	ٻنھي	-	-	
		tʰalhə (platter)	ʉnʰi (both)			
2.6	/ə/, /əʰ/, short vowel, or no sound at all	نہ	-	-	-	
		nə (no/don't)				

Having looked at the desired shapes, it is now important to compare them to the existing *heh* characters in Unicode.

Information in the Unicode Standard regarding *heh* and *heh doachashmee*

The issue of which HEH to use in Sindhi is not discussed in the Unicode chapter on Arabic. However, this information is on pages 380-381:

Letter heh. In the case of U+0647 ARABIC LETTER HEH, the glyph ه is shown in the code charts. This form is often used to reduce the chance of misidentifying HEH as U+0665 ARABIC-INDIC DIGIT FIVE, which has a very similar shape. The isolated forms of U+0647 ARABIC LETTER HEH and U+06C1 ARABIC LETTER HEH GOAL both look like U+06D5 ARABIC LETTER AE.

U+06BE ARABIC LETTER HEH DOACHASHMEE is used to represent any heh-like letter that appears with stems at both sides in all contextual forms. The exact contextual shapes of the letter depend on the language and the style of writing. The forms shown in Table 9-8 for KNOTTED HEH are used in certain styles of writing in South Asia. Other South Asian styles may use different medial and final forms. The style used in China and Central Asia for languages such as Uyghur uses medial and final forms for HEH DOACHASHMEE that are visually similar to the medial form of HEH shown in Table 9-8.

Table 9-8 includes the following glyphs for U+0647 HEH and U+06BE HEH DOACHASHMEE

HEH	ه	ه	ه	ه
KNOTTED HEH	ه	ه	ه	ه

The table below also takes into account the compatibility block:

USV	Source for shapes	Final	Medial	Initial	Isolate	All forms
Dual-joining						
U+0647 HEH	from Arabic compatibility block	ه	ه	ه	ه	ه ه ه
U+06BE HEH DOACHASHMEE	common shape	ه	ه	ه	ه	ه ه ه
	from Arabic compatibility block	ه	ه	ه	ه	ه ه ه
SINDHI HEH		ه	ه	ه	ه	ه ه ه
SINDHI ASPIRATED HEH		ه	ه	-	-	ه ه -

The “knotted heh”, or *heh doachashmee* from the *Arabic compatibility block* is visually most similar to the Sindhi *heh* (see blue). However, this graphic does not show the complete picture. U+06BE is generally used for aspirated consonants, and the Sindhi *heh* is used as a normal *heh*, not for aspiration. Additionally, the SINDHI ASPIRATED HEH is visually similar to the *common shape* for *heh doachashmee* (see green)! Another important aspect is that most *nastaliq* fonts use the *compatibility shape*, and most *naskh* fonts use the *common shape*.

There are a few figures below in the “Samples” section. However, most Sindhi books, newspapers, and images that were found made inconsistent use of these shapes. However, the Sindhi Language Authority has made it clear which shapes are required [2].

Options

Four possibilities for a solution are considered here. These would be to:

1. consider the Sindhi *heh* a glyph variant of U+0647
 - a. the Sindhi *aspirated heh* should use U+06BE
 2. use U+06BE for the Sindhi *heh*
 - a. consider the Sindhi *aspirated heh* a glyph variant of U+0647
 3. use U+06BE for the Sindhi *heh*
 - a. the Sindhi *aspirated heh* would use U+0647 **and** U+06C1 (if word final)
 4. use U+06BE for the Sindhi *heh*
 - a. encode a new Sindhi *aspirated heh*
1. Consider the Sindhi *heh* a glyph variant of U+0647 HEH and *aspirated heh* should use U+06BE

See figure 4.

This solution would treat the Sindhi *heh* as a variant of U+0647. This solution is documented in Kew, section 3.6, page 7 [2]. The Sindhi *heh* represents the same sound as it does in the Arabic language. No change would be needed for the *aspirated heh*, since this is the **common shape** for the *heh doachashmee* (since the *aspirated heh* is always preceded by another character, the isolate and initial forms are not needed).

	Default / Common Shape	Sindhi Shape
U+0647		
U+06BE		

This solution would be attractive to SIL users, because this is the solution implemented in SIL fonts for about 20 years. Existing documents would not need converting to a new codepoint. Since the *sound* for the character is equivalent to the normal usage for U+0647, it makes perfect sense to use U+0647 as the codepoint.

Outcome with this option:

Chapter 9 in the Core Spec should be updated to give an image of the Sindhi *heh*. Suggested additional text in blue below:

Letter heh. In the case of U+0647 ARABIC LETTER HEH, the glyph ه is shown in the code charts. This form is often used to reduce the chance of misidentifying heh as U+0665 ARABIC-INDIC DIGIT FIVE, which has a very similar shape. The isolated forms of U+0647 ARABIC LETTER HEH AND U+06C1 ARABIC LETTER HEH GOAL both look like U+06D5 ARABIC LETTER AE. **However, for Sindhi the isolate, initial, medial, and final forms are knotted.**

Table 9.x Sindhi HEH design

Common Shape	Sindhi Shape

No properties need adjusting for these characters. Since we are not asking for a change in the charts, the Presentation forms block should not need adjusting.

There would be no documentation changes required for U+06BE since this is the **common shape** for U+06BE.

This solution does not take into account the rare requirement for the standard *heh* used in borrowed words.

2. Use U+06BE HEH DOACHASHMEE for Sindhi *heh* and consider the *aspirated heh* a variant of U+0647 HEH

In this solution the Sindhi *heh* would be encoded as U+06BE and it would use the glyph shapes from the Arabic compatibility block. The *aspirated heh* would be considered a variant of U+0647 HEH where the final form is a variant (since the *aspirated heh* is always preceded by another character, the isolate and initial forms are not needed).

	Compatibility Shape	Sindhi Shape
U+06BE		
U+0647		

Outcome with this option:

There would be no documentation changes required for U+06BE since this is the *common shape* for the *heh doachashme*. Chapter 9 in the Core Spec should be updated to give an image of the Sindhi *heh*.

Table 9.x Sindhi heh design

Common Shape	Sindhi Shape
	

No properties need adjusting for these characters. Since we are not asking for a change in the charts, the Presentation forms block should not need adjusting.

This solution does not take into account the rare requirement for the standard *heh* used in borrowed words.

3. Use U+06BE HEH DOACHASHMEE for Sindhi heh and the Sindhi aspirated heh would use U+0647 HEH and U+06C1 (if word final)

See figure 5.

In this solution the Sindhi *heh* would be encoded as U+06BE and it would use the glyph shapes from the Arabic compatibility block. The *aspirated heh* would be encoded as U+0647 HEH. In the case where the *aspirated heh* is word final, a U+06C1 would be inserted.

This solution would be attractive to the Sindhi Language Authority (SLA) as this has been implemented in version 2.0 of the SLA fonts and keyboard.

Searching for words and letters would be somewhat more difficult as U+06C1 would need to be included in searches and sorting when looking for aspirated words.

	Compatibility Shape	Sindhi Shape
U+06BE		
U+0647		

This solution allows for use of the standard *heh* in borrowed words.

Outcome with this option:

This solution would require fonts to use the compatibility shapes for U+06BE.

Input routines would need to be adjusted so that the user could insert U+06C1 at the end of aspirated words.

In this solution, not much would change on the Unicode side of things. However, it would be important to document the use of U+06C1 HEH GOAL with HEH in the core specification.

Feedback on this option:

The author is grateful to Jonathan Kew for a review of this document. Kew was quite between 2001-2003 in encoding Arabic script characters into Unicode. He was either the author of all or most of the proposals for the characters encoded in the **Arabic Supplement** block. These are his comments.

It's my suspicion that this idea derives from the world of legacy 8-bit font encodings (possibly based on earlier lead-type practice), where a very limited selection of glyphs were available and extending the set was difficult and/or expensive. People found that Arabic fonts had a "doachashmee" shape as the medial form of /*heh*/, but if this glyph was used at the end of a word it would look abruptly truncated (because it was designed to be a medial glyph).

When at the end of a word, it needed some kind of "tail", an end that tapers off nicely rather than just being cut off. Creating an entire new glyph for "final heh-doachashmee" was out of reach for would-be writers of Sindhi. But the final shape of "heh goal", which had been created for Urdu use, provided a reasonable solution: it's just a little tail that neatly finishes the word without leaving the abrupt end of the medial doachashmee glyph.

(I've also seen occasional examples where an author has used a trailing kashida character rather than an extra heh to force the aspiration-heh into a doachashmee shape; and sometimes it appears simply cut off, as if followed by ZWJ although in fact more likely the product of a purely glyph-based system where no automatic shaping is happening.)

So I think this convention, where /*jh*/ and /*gh*/ (which are regarded as separate letters of the Sindhi alphabet, although written as digraphs rather than the single characters of other aspirates like /*ph*/, /*bh*/, /*th*/, etc) become *trigraphs* at the end of a word is perpetuating a glyph-based encoding hack that came about when people were trying to shoehorn Sindhi into systems that were created to support Arabic and Urdu. Looking at a Sindhi alphabet

chart, it seems clear to me that conceptually */jh/* and */gh/* are composed of two elements, a root consonant */j/* or */g/* plus a doachashmee-shaped */h/* representing the added aspiration.

Users faced with systems where the only doachashmee-shaped glyph was a medial */heh/* have worked around this, most often by adding a second */heh/* when the aspirated consonant appears in final (or isolated) position, but it would be unfortunate to enshrine this technical workaround as a permanent “spelling” rule of the writing system when encoded in Unicode.

4. Disunify U+06BE HEH DOACHASHMEE common shape and compatibility shape

Since shapes of glyphs can change, but compatibility decompositions cannot change, it would be best to keep U+06BE ARABIC LETTER HEH DOACHASHMEE with the *compatibility shape* (هه ه) to match the compatibility decompositions of U+FBAA..U+FBAD.

Then, a new character (possibly called ARABIC LETTER ALTERNATE HEH DOACHASHMEE) could be encoded that uses the *common shape* (هه ه) for U+06BE HEH DOACHASHMEE. Sindhi only requires the medial and final forms, but encoding a new character with only medial and final forms would introduce an unwelcome new aspect to the Arabic encoding model (not dual-joining which has four forms, nor right-joining which only has isolate and final forms).

This third option would be an attractive solution because the shapes for the Sindhi *heh* are the same as the compatibility block shapes for *heh doachashmee*.

	Compatibility block	Sindhi
U+06BE	هه ه	هه ه
		هه ه

Outcome with this option:

U+06BE would use the *compatibility shape* (هه ه) for the *heh doachashmee*.

A new *aspirated heh* (هه ه) would be encoded with the same properties as *heh doachashmee*.

Multiple system fonts would need modifying (Amiri, Arabic Typesetting, Arial, Calibri, Courier New, DejaVu Sans, Microsoft Sans Serif, Microsoft Uighur, Noto Sans Arabic (but not other Noto Arabic fonts), Sakkal Majalla, Segoe, Simplified Arabic, Tahoma, Times New Roman, Urdu Naskh Unicode).

Unicode character Properties

```
08xx; ARABIC LETTER ALTERNATE HEH DOACHASHMEE;Lo;0;AL;;;;;N;;;;;
```

Joining type and group for ArabicShaping.txt

```
08xx; ALTERNATE HEH DOACHASHMEE; D; ALTERNATE HEH DOACHASHMEE
```

Normalization and Confusability Issues

Since the *sindhi heh* is not considered the same character as U+0647 or U+06BE, there should not be any decomposition nor normalization issues.

ARABIC LETTER ALTERNATE HEH DOACHASHMEE might be confusable with U+06BE.

Suggested Collation

Suggested collation is for the ALTERNATE HEH DOACHASHMEE to come after HEH DOACHASHMEE (U+06BE).

Core Specification

Chapter 9 in the Core Spec should be updated. Suggested additional text in blue below:

U+06BE ARABIC LETTER HEH DOACHASHMEE is used to represent any heh-like letter that appears with stems at both sides in all contextual forms. The exact contextual shapes of the letter depend on the language and the style of writing. The forms shown in *Table 9-8* for KNOTTED HEH are used in certain styles of writing in South Asia. Other South Asian styles may use different medial and final forms. The style used in China and Central Asia for languages such as Uyghur uses medial and final forms for HEH DOACHASHMEE that are visually similar to the medial form of HEH shown in *Table 9-8*. [U+08xx ARABIC LETTER ALTERNATE HEH DOACHASHMEE was added to Unicode to support the Sindhi language. This is shown in *Table 9-8*.](#)

“Table 9-8. Dual-Joining Arabic Characters” would need to be updated to include the ALTERNATE HEH DOACHASHMEE shapes and shaping group.

Samples

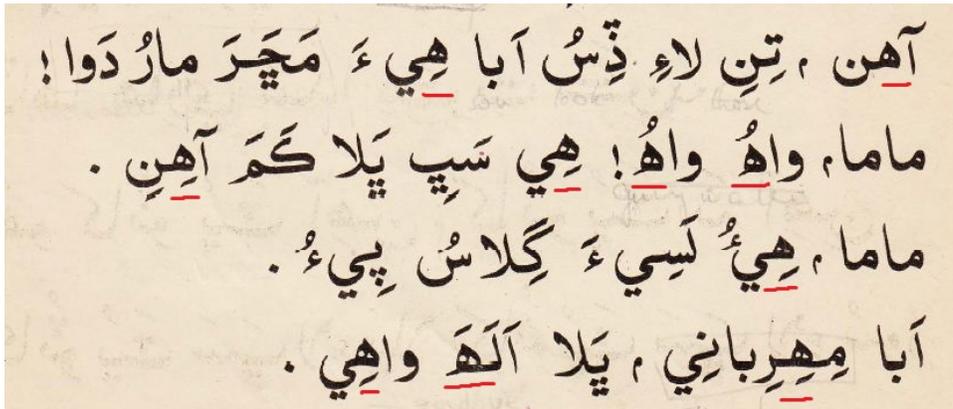


Figure 1: Sindhi heh (red) isolate, initial, medial and final forms [5, page 12]

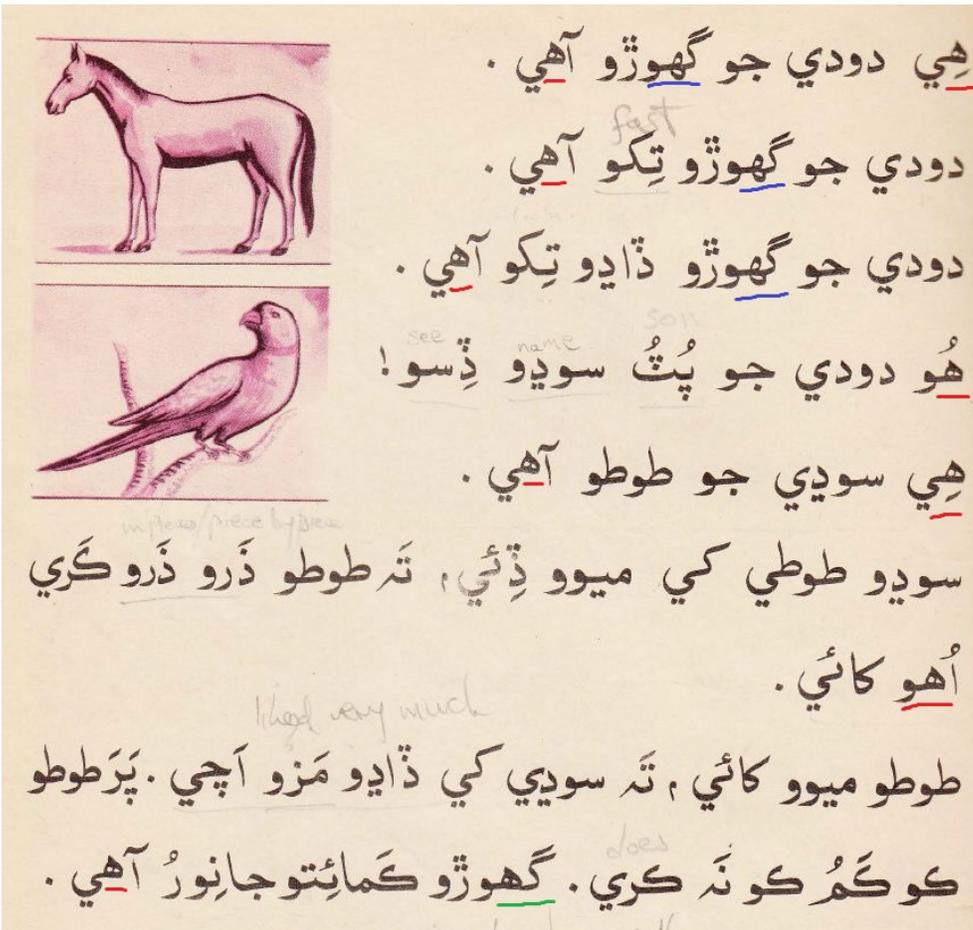


Figure 2: Sindhi heh (red), initial form. Sindhi aspirated heh (blue), medial form. Sindhi aspirated heh (green) inconsistent shape for aspirated heh. [5, page 17]

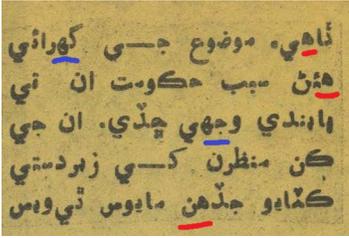


Figure 3: Sindhi heh (red), initial form. Sindhi aspirated heh (blue), medial form. [4, back page]

Transl.	Sindhi in Arabic script			
	Final	Medial	Initial	Isolate
/jh/ (U+06BE)	جھ	جھ	جھ	جھ
/gh/ (U+06BE)	گھ	گھ	گھ	گھ
/h/ (U+0647)	ھ or ھ	ھ or ھ	ھ	ھ

Figure 4a: Sindhi heh codepoint recommendations from Kew [1, section 3.4]

<i>Typical default shapes for...</i>		<i>Isolate</i>	<i>Final</i>	<i>Medial</i>	<i>Initial</i>
U+0647	ARABIC LETTER HEH	ه	ه	ه or ه	ه
U+06BE	... DOACHASHMEE	ھ	ھ	ھ	ھ
U+06C1	... GOAL	ه	ه	ه	ه
U+06FF	... WITH INVERTED SMALL V ABOVE	ھ	ھ	ھ	ھ
<i>Urdu</i>					
U+0647	ARABIC LETTER HEH	ه	ه	ه	ه
<i>Sindhi</i>					
U+0647	ARABIC LETTER HEH	ھ	ھ or ھ	ھ or ھ	ھ
<i>Parkari</i>					
U+0647	ARABIC LETTER HEH	ھ	ھ	ھ	ھ
<i>Kurdish</i>					
U+0647	ARABIC LETTER HEH	ھ	ھ	ھ	ھ

Figure 4b: recommended default shapes for various languages from Kew [1, section 5]

MB Khursheed SK 2.0

فانت ۾ ”ه“ جي شڪلين جو لکت ۾ واهپو

ه ه ه ه ه ه ه

توڪي ڪهڙي ڪل نه گهر جي چُله ڪيئن ٿي پري؟
تون ته باچاه آهين، هميشه وٽين آواره گري ڪندو،
اَسَر جو نڪرين ته سَنجها مهل پيو موٽين.
ميان! جُلَم ڪرڻ ڏاڍي سولي، پر ٻوٽو ٻارڻ ڏاڍو ڏکيو آ.

MB Sookhri SK 2.0

فانت ۾ ”ه“ جي شڪلين جو لکت ۾ واهپو

ه ه ه ه ه ه ه

توڪي ڪهڙي ڪل ته گهر جي چُله ڪيئن ٿي پري؟
تون ته باچاه آهين، هميشه وٽين آواره گري ڪندو،
اَسَر جو نڪرين ته سَنجها مهل پيو موٽين.
ميان! جُلَم ڪرڻ ڏاڍي سولي، پر ٻوٽو ٻارڻ ڏاڍو ڏکيو آ.

MB Lateefi SK 2.0

فانت ۾ ”ه“ جي شڪلين جو لکت ۾ واهپو

ه ه ه ه ه ه ه

توڪي ڪهڙي ڪل ته گهر جي چُله ڪيئن ٿي پري؟
تون ته باچاه آهين، هميشه وٽين آواره گري ڪندو،
اَسَر جو نڪرين ته سَنجها مهل پيو موٽين.
ميان! جُلَم ڪرڻ ڏاڍي سولي، پر ٻوٽو ٻارڻ ڏاڍو ڏکيو آ.

Figure 5a: Sindhi heh shapes from SLA, [2]

	چُله	گهر	ڪهڙي
	U+0686 U+064F U+0644 U+0647 U+06C1	U+06AF U+0647 U+0631	U+06AA U+06BE U+0699 U+064A
آواره	هميشه	آهين	باچاه
U+0622 U+0648 U+0627 U+0631 U+06C1	U+06BE U+0645 U+064A U+0634 U+0647	U+0622 U+06BE U+064A U+0646	U+0628 U+0627 U+0687 U+0627 U+06BE
		مهل	سَنجها
		U+0645 U+06BE U+0644	U+0633 U+064E U+0646 U+062C U+0647 U+0627
			جُلَم
			U+062C U+064F U+0644 U+0650 U+06BE U+0650

Figure 5b: Sindhi heh shapes from SLA (with words and codepoints), [2]

References

- [1] Kew, Jonathan. 2005. [Notes on some Unicode Arabic characters: recommendations for usage \(draft 2\)](#). SIL International. (Accessed 24 March 2021)
- [2] Kumbhar, Shabir. [سنڌي لکت لاءِ ”ه“ جون گھربل شڪليون](#). Required Forms of Heh (for Writing Sindhi). (Accessed 24 March 2021)
- [3] Pournader, Roozbeh. 2014. [The right hehs for Arabic script orthographies of Sorani Kurdish and Uighur](#). (Accessed 22 July 2020) Minutes: <https://www.unicode.org/L2/L2014/14100.htm> Still an open action item for Roozbeh: <https://www.unicode.org/L2/L-SD2.htm>
- [4] Sindhi Newspaper.
- [5] Sindhi Primer for Adults. 1973.
- [6] Unicode. [Chapter 9 Middle East-I: Modern and Liturgical Scripts](#). (accessed 23 March 2021).
- [7] Unicode FAQ. Middle Eastern Scripts and Languages, <http://www.unicode.org/faq/middleeast.html> (accessed 29 July 2020).
- [8] Wikipedia. Urdu Alphabet https://en.wikipedia.org/wiki/Urdu_alphabet (accessed 26 August 2020).
- [9] [MBSindhi MSKLC](#) keyboard (outdated and possibly has malware).
- [10] April 2018, an update of [MB Sindhi SK 2.0 Keyboard pack](#) was released in which all forms of U+06BE and isolate U+0647 were corrected in the 24 most useable [standard Sindhi fonts](#).
- [11] Sindhi Language Authority (SLA). https://en.wikipedia.org/wiki/Sindhi_Language_Authority (accessed 29 July 2020).
- [12] Noto Naskh Arabic fonts. <https://noto-website-2.storage.googleapis.com/pkgs/NotoNaskhArabic-hinted.zip> (accessed 26 August 2020).
- [13] Noto Sans Arabic fonts. <https://noto-website-2.storage.googleapis.com/pkgs/NotoSansArabic-hinted.zip> (accessed 26 August 2020).