

Sindhi has its own set of numerals based on numerals used in Arabic, Persian and Urdu. These numerals are listed in Figure 7.4.

۹ ۸ ۷ ۶ ۵ ۴ ۳ ۲ ۱ ۰

Figure 7.4. Sindhi Numerals

7.1.2. Bidirectionality

Sindhi inherits the bidirectional property from Arabic script. Sindhi words are written from right to left but numbers are written from right to left, as shown in Figure 7.5. However, bidirectionality is handled at rendering level and key press sequence for Sindhi alphanumeric input is same as it would be for any other uni-directional language. Thus bidirectionality has no implication on collation.

سنڌي ۱۲۳ بلاگ

Figure 7.5. Bidirectional Sindhi Text

(Arrows indicate reading direction)

7.1.3. Cursiveness, Ligation and Context Sensitive Glyph Shaping

Arabic script is cursive, that is, the letters in the script join together into units to form words. These connected units are called ligatures. There are two kinds of characters, joiners and non-joiners. While writing a word, all characters join together until a non-joiner is written. A new ligature starts after the non-joiner (thus, the name “non-joiner”). The process is repeated until the end of the word. In addition, depending on whether the character joins a ligature in the initial, medial or final position, or is unconnected, it takes a different shape. Cursiveness is shown in Figure 7.6.

سنڌي
Cursively Written Form
س ن ڌ ي
Spelling

Figure 7.6. Spelt-out and Cursive Version of Sample Text of Sindhi

Again, cursiveness, ligation and context sensitivity are rendering related issues and the though the output shapes of characters may vary with context, their internal encoding remains

unchanged. For example, the letter ب may take multiple shapes but its internal encoding is always U+0628. Therefore, these properties have no implication on collation.

7.2. Collation

Sindhi collation sequence has been standardized and published by Sindhi Language Authority for Pakistan. The collation requires the characters to be sorted at three levels, letters, Aerab and honorifics. However, before the text can be sorted, it has to undergo text processing, as discussed in the next sub-section. Once the text is processed and collation elements are assigned, the regular sort-key generation and comparison process sorts the text.

7.2.1. Text Processing

7.2.1.1. Inconsistent Use of Space

Naskh style of writing does not have a strong concept of space to separate words. Similar to South-East Asian scripts like Lao, Thai and Khmer, Sindhi readers are expected to parse the ligatures into words as they read along the text. This has implications on collation and thus proper word segmentation must be done before strings are collated. Currently there are no automatic word segmentation utilities available for Sindhi and therefore the input for collation must be manually cleaned.

7.2.1.2. Normalization

Two kinds of normalization are required for Sindhi. First, a letter may be represented by multiple Unicode points, and thus the redundancy in encoding has to be cleaned in raw text before further processing. For example, letter ج may be represented by Unicode points U+0649, U+064A, and

U+06CC in Sindhi. Second, a letter or a ligature is sometimes encoded in composed form as well as decomposed form. Thus, the two equivalent representations must also be reduced to same underlying form before further processing. Table 7.1 below gives an example.

Table 7.1. Composed and Decomposed Forms of a Sindhi Ligature

Ligature Glyph	Unicode	Individual letters/marks	Unicode Points
ج	FEFB	ا ج	0627 06F1

There are many such ligatures which can be represented in multiple ways. Many are not recommended by the Unicode standard, but users still use them due to the similarity of glyphs. An example is using Arabic digits for Sindhi language (U+0660 – U+0669), where a separate similar looking set is also encoded (U+06F0 – U+06F9) for use of Arabic language.

7.2.1.3. Contraction

In Sindhi character ھ (U+06BE or U+0647¹) combines with two letters ج and گ to represent their aspirated versions. Though the constituents are encoded separately, they combine to give a singular character with a single collation element. Thus, these combinations have to be contracted before collation elements are assigned. Some examples of these contractions are given in Figure 7.7.

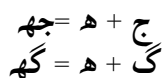


Figure 7.7. Contraction of Letters with ھ in Sindhi

There is no Unicode point available to directly encode the contracted form for the aspirated versions shown in the figure.

7.2.2. Unicode Collation Elements

Collation Elements for Sindhi character set are given in Table 7.2 below. These are based on [44]. Also see [6] for additional background information.

Table 7.2. Sindhi Collation Elements

Glyph	Unicode	Collation Elements	Unicode Name
← Numerals →			
٠	06F0	0E29 0020 0002	ARABIC-INDIC DIGIT ZERO
١	06F1	0E2A 0020 0002	ARABIC-INDIC DIGIT ONE
٢	06F2	0E2B 0020 0002	ARABIC-INDIC DIGIT TWO
٣	06F3	0E2C 0020 0002	ARABIC-INDIC DIGIT THREE
٤	06F4	0E2D 0020 0002	ARABIC-INDIC DIGIT FOUR
٥	06F5	0E2E 0020 0002	ARABIC-INDIC DIGIT FIVE
٦	06F6	0E2F 0020 0002	ARABIC-INDIC DIGIT SIX
٧	06F7	0E30 0020 0002	ARABIC-INDIC DIGIT SEVEN
٨	06F8	0E31 0020 0002	ARABIC-INDIC DIGIT EIGHT

¹ Not recommended for use for Sindhi.

٩	06F9	0E32 0020 0002	ARABIC-INDIC DIGIT NINE
← Consonants and Vowels →			
ا	0627	1350 0020 0002	ARABIC LETTER ALEF
ب	0628	1353 0020 0002	ARABIC LETTER BEH
بٲ	067B	1356 0020 0002	ARABIC LETTER BEEH
بٲٲ	0680	1359 0020 0002	ARABIC LETTER BEHEH
ت	062A	135C 0020 0002	ARABIC LETTER TEH
تٲ	067F	135F 0020 0002	ARABIC LETTER TEHEH
تٲٲ	067D	1360 0020 0002	ARABIC LETTER THE WITH THREE DOTS ABOVE DOWNWARDS
تٲٲٲ	067A	1363 0020 0002	ARABIC LETTER TTEHEH
ث	062B	1366 0020 0002	ARABIC LETTER THEH
پ	067E	1369 0020 0002	ARABIC LETTER PEH
پٲ	06A6	136C 0020 0002	ARABIC LETTER PEHEH
ج	062C	136F 0020 0002	ARABIC LETTER JEEM
جٲ	0684	1370 0020 0002	ARABIC LETTER DYEH
جٲٲ	062C 06BE	1373 0020 0002	ARABIC LETTER JEEM + ARABIC LETTER HEH DOCHASHMEE
جٲٲٲ	0683	1376 0020 0002	ARABIC LETTER NYEH
چ	0686	1379 0020 0002	ARABIC LETTER TCHEH
چٲ	0687	137C 0020 0002	ARABIC LETTER TCHEHEH
ح	062D	137F 0020 0002	ARABIC LETTER HAH
خ	062E	1380 0020 0002	ARABIC LETTER KHAH
د	062F	1383 0020 0002	ARABIC LETTER DAL
دٲ	068C	1386 0020 0002	ARABIC LETTER DAHAL
دٲٲ	068F	1389 0020 0002	ARABIC LETTER DAL WITH THREE DOTS ABOVE DOWNWARD
دٲٲٲ	068A	138C 0020 0002	ARABIC LETTER DAL WITH DOT BELOW
ڍ	068D	138F 0020 0002	ARABIC LETTER DDAHAL
ذ	0630	1390 0020 0002	ARABIC LETTER THAL
ر	0631	1393 0020 0002	ARABIC LETTER REH
رٲ	0699	1396 0020 0002	ARABIC LETTER REH WITH FOUR DOTS ABOVE
ز	0632	1399 0020 0002	ARABIC LETTER ZAIN
س	0633	139C 0020 0002	ARABIC LETTER SEEN
ش	0634	139F 0020 0002	ARABIC LETTER SHEEN
ص	0635	13A0 0020 0002	ARABIC LETTER SAD
صٲ	0636	13A3 0020 0002	ARABIC LETTER DAD
ط	0637	13A6 0020 0002	ARABIC LETTER TAH
ظ	0638	13A9 0020 0002	ARABIC LETTER ZAH

ع	0639	13AC 0020 0002	ARABIC LETTER AIN
غ	063A	13AF 0020 0002	ARABIC LETTER GHAIN
ف	0641	13B0 0020 0002	ARABIC LETTER FEH
ق	0642	13B3 0020 0002	ARABIC LETTER QAF
ك	06AA	13B6 0020 0002	ARABIC LETTER SWASH KAF
ك	06A9	13B9 0020 0002	ARABIC LETTER KEHEH
گ	06AF	13BC 0020 0002	ARABIC LETTER GAF
گھ	06B3	13BF 0020 0002	ARABIC LETTER GUEH
گھ	06AF 06BE	13C0 0020 0002	ARABIC LETTER GAF + ARABIC LETTER HEH DOCHASHMEE
نگھ	06B1	13C3 0020 0002	ARABIC LETTER NGOEH
ل	0644	13C6 0020 0002	ARABIC LETTER LAM
م	0645	13C9 0020 0002	ARABIC LETTER MEEM
ن	0646	13CC 0020 0002	ARABIC LETTER NOON
نٹ	06BB	13CF 0020 0002	ARABIC LETTER RNOON
و	0648	13D0 0020 0002	ARABIC LETTER WAW
ہ	06C1	13D3 0020 0002	ARABIC LETTER HEH GOAL
ھ	06BE	13D6 0020 0002	ARABIC LETTER HEH DOCHASHMEE
ء	0621	13D9 0020 0002	ARABIC LETTER HAMZA
ی	06CC	13DC 0020 0002	ARABIC LETTER FARSI YEH
← Diacritics →			
◌ْ	0652	0000 00C4 0002	ARABIC SUKUN
◌َ	064E	0000 00C9 0002	ARABIC FATHA
◌ِ	0650	0000 00CA 0002	ARABIC KASRA
◌◌	064F	0000 00CB 0002	ARABIC DAMMA
◌ ^۰	0670	0000 00CD 0002	ARABIC LETTER SUPERScript ALEF
◌◌◌	0651	0000 00E8 0002	ARABIC SHADDA
← Honorifics and Special Signs →			
◌◌◌◌	0610	0000 0000 000A	ARABIC SIGN SALLALLAHOU ALAYHWASSALLAM
◌◌◌◌	0611	0000 0000 001A	ARABIC SIGN ALAYHE ASSALLAM
◌◌◌◌	0613	0000 0000 002A	ARABIC SIGN RADI ALLAHOU ANHU
◌◌◌◌	0612	0000 0000 003A	ARABIC SIGN RAHMATULLAH ALAYHE
← Punctuation Marks (Ignorable) →			
◌◌◌	0615	0000 0000 0000	ARABIC SMALL HIGH TAH

،	060C	0000 0000 0000	ARABIC COMMA
،	060D	0000 0000 0000	ARABIC DATE SEPARATOR
٫	066B	0000 0000 0000	ARABIC DECIMAL SEPARATOR
،	066C	0000 0000 0000	ARABIC THOUSANDS SEPARATOR
؟	061F	0000 0000 0000	ARABIC QUESTION MARK
؛	061B	0000 0000 0000	ARABIC SEMICOLON
-	06D4	0000 0000 0000	ARABIC FULL STOP
%	066A	0000 0000 0000	ARABIC PERCENT SIGN
لا	FEFB	[13AB 0020 0002],[1350 0020 0002]	ARABIC LIGATURE LAAM WITH ALEF ISOLATED FORM
الله	FDF2	[13AB 0020 0002], [13AB 0020 0002], [13AB 0020 0002],[13D3 0020 0002]	ARABIC LIGATURE ALLAH

Results

The sorting performed using the collation elements given results in the following sequence.

Table 7.3. Input and Corresponding Sorted Output for Sindhi

Sample Output		Sample Input	
ڊر	آرياکرڻ	ڦٽڪو	ڪاراوڻ
رائيڻ	آريڪڻ	ڦٽڪي	شهوت
سفر	آڙچڻ	ڦٽائڻ	صبر
سڦرو	ڦٽ	ڦٽائڻ	ضيڪ
سڦرو	ڦٽ	ڦٽو	طوفان
شهوت	ڦٽ	ڦٽو	عظميٰ
صبر	ڦٽائڻ	ڦٽڙائڻ	قشيميش
ضيڪ	ڦٽڪو	ڦٽڪڻ	ڪامول
طوفان	ڦٽڪي	ڦٽو	ڪاتو
عظميٰ	ڦٽائڻ	ڦٽو	ڪابو
قشيميش	ڦٽو	ڦٽو	آريڪڻ
ڪاراوڻ	ڦٽو	ڦٽو	گنڪا
ڪامول	ڦٽڙائڻ	ڦٽو	گهلاڻو
ڪابو	ڦٽڪڻ	ڦٽو	گهر
ڪاتو	ڦٽو	ڦٽو	لاڪڙ
گنڪا	ڦٽو	ڦٽو	لڳن

گهر گهلاڻو لاکڙ لگن منارڻ وات هت يتيم	ڏوپ ڏوپ ڏوپ ڏاه ڏاه ڏاهي ڏپ ڏپ ڏپر	ڏپ ڏپ ڏپر ڏر رانيڻ سفر سڦرو سڦرو	منارڻ وات هت يتيم آرياڪرڻ آڙچڻ ڦٽ ڦٽ ڦٽ
--	--	---	---

7.3. Conclusion

Sorting in Sindhi is carried out at three different levels. Letters are sorted at primary level, diacritics are handled at secondary level, and honorifics are handled at tertiary level. Normalization and contraction are also required for Sindhi collation. However, regular sorting algorithm is applicable after appropriate text processing is done and collation elements are assigned.