

Table 8.1. Special Marks in Sinhala

Name	Glyph	Usage
Virama / Al-Lakuna	◌̣	ක̣
Anusvara	◌◌̣	ක◌̣
Visarga	◌◌̣◌̣	ක◌̣◌̣

Al-Lakuna is discussed in the section below. The Anusvara and Visarga are semi-consonants and can occur only with vowels [48]. Anusvara is used for nasalization and also to indicate the actual [n] sound at the end of a syllable [5]. Visarga is used for aspiration of vowels.

Sinhala does not have its own set of numerals and 0, 1, 2... 9 are used.

8.1.2. Script Details

Sinhala is written from left to right. Letters are uncased and are grouped based on their place and manner of articulation, like other Indic scripts. Traditionally space was not used, and only a special punctuation mark Kunddaliya (කුඳ්දලියා U+0DF4) was employed at the end of paragraph. Now spaces are used with European punctuation [5].

8.1.2.1. Consonants and Vowels

Sinhala has a syllabic writing system like other Indic based languages. Vowels and consonants are not represented as an individual unit like Latin script rather as syllabic units in which consonant has an inherent [a/ə] vowel, if not otherwise specified. For example Sinhala Letter Alpapraana kayanna ක has [ka/kə] sound. In case the consonant is to be articulated without a vowel sound, e.g. in a cluster or at the end of a word, Al-Lakuna is placed at top right of the consonant to cancel the [a] sound. So ක̣ has [k] sound.

Independent vowels are used for syllables which do not have an onset consonant and thus start with a vowel. For all syllables which have an onset consonant, dependent vowels attach with this consonant. If the consonant is followed by a vocalic sound different from [a], the appropriate dependent vowel mark is attached before, after, above or below the consonant (though it always logically follows the consonant). In some cases the vowel splits into two halves and is placed around the consonant. Table 8.2 below shows these cases.

Table 8.2. Dependent Vowels with the Consonant [k]

Consonant + Dependent Vowel	Joined Form	Comment
ක + ඌ	කෙ	Connects to the left of consonant
ක + ඌ	කෑ	Connects to the right of consonant
ක + ඌ	කී	Connects to the top of consonant
ක + ඌ	කු	Connects to the base of consonant
ක + ඌ	කො	Wraps around the consonant

Only one vowel can occur in a syllable, thus only a single dependent vowel can attach with a consonant and the dependent vowels can not occur with independent vowels.

8.1.2.2. Conjunct Consonants and Consonantal Vowel Ligatures

Sinhala also forms conjunct consonants (known as *bændi akuru*). Unique combined shapes or ligatures are formed when characters ර (or 'ra') and ය (or 'ya') combine with consonants or when other consonants form a cluster within a syllable [51]. Two such examples are given in Table 8.3.

Table 8.3. Conjuncts in Sinhala [51]

Individual Letters	Conjoined Form
ක + ඌ + ර	කුර
ක + ඌ + ය	කුය

8.2. Collation

Sinhala collation sequence, as followed by the dictionaries, is being standardized through Sri Lankan authorities, and draft is already in consideration. This section elaborates on this collation sequence for Sinhala and an algorithmic implementation using UCA [2].

In Sinhala all characters have primary level significance for collation purposes. The relative order is also well defined: vowels, then semi-consonants and finally consonants [48]. However, before collation can be applied, some text processing is required. These details are also given below.

8.2.1. Text Processing

8.2.1.1. Reordering

As shown in Table 8.2 above, independent vowels combine with consonants in different ways. In hand-written orthography, old type-writers and proprietary single byte Sinhala fonts, the vowels that append to the left are written first followed by a consonant. The vowels that append to the right, above or below are written after the consonant. However, the logical order in both cases is the same, i.e. the consonant is followed by the vowel. The more recent font formats and fonts follow the logical order of typing. However, for the legacy fonts, re-ordering will be required before string comparisons can be performed for collation.

8.2.1.2. Normalization

Many Sinhala vowels are formed with two parts, one part attaching before and other after the following consonant. These and some other dependent forms of vowels can be encoded in more than one way in Unicode. As they are equivalent to each other for text processing, they have to be equated or normalized into the same composed or decomposed form. Some examples are illustrated in Table 8.4 below.

Table 8.4. Normalization in Sinhala

Decomposed Form	Unicode of Decomposed Form	Equivalent Composed Form	Unicode of Composed Form
ඉ ් ධ	0DDC + 0DCA	ඉධ	0DDD
ඉ ් ඉ	0DD9 + 0DCF	ඉඉ	0DDC
ඉ ් ඉ	0DD9 + 0DDF	ඉඉ	0DDE

8.2.1.3. Contraction

In case the encoding is being translated into decomposed form, contraction is needed for assigning the collation elements, i.e. multiple character codes would map onto a single collation element. This contraction for consonants and vowels is illustrated in Table 8.5.

Table 8.5. Contraction to Single Collation Element from Multiple Unicodes

Glyph	Unicodes of Decomposed Form	Unicode of Composed Form	Collation Element	Unicode Name
෧෦ ෧෦ = ෧෦෦	0DD8 0DD8	0DF2	1410 0020 0002	SINHALA VOWEL SIGN DIGA GAETTA- PILLA
෧෧ ෧෦ = ෧෦෦	0DD9 0DCA	0DDA	141A 0020 0002	SINHALA VOWEL SIGN DIGA KOMBUVA
෧෧෧ ෧෦ = ෧෦෦෦	0DDC DCA	0DDD	1420 0020 0002	SINHALA VOWEL SIGN KOMBUVA HAA DIGA AELA-PILLA

8.2.1.4. Conjuncts

The formation of conjuncts causes visual changes but does not change input sequence logically. Therefore it has no bearing on the collation process.

8.2.2. Unicode Collation Elements

In order to realize Sinhala collation the following collation elements need to be assigned. The UCA algorithm proposed in [2] may be applied for sorting. The realized sequence is same as recommended by [48, 49].

Table 8.6. Sinhala Collation Elements

Glyph	Unicode	Collation Elements	Unicode Name
←Independent Vowels →			
අ	0D85	1356 0020 0002	SINHALA LETTER AYANNA
ආ	0D86	1359 0020 0002	SINHALA LETTER AAYANNA
ඇ	0D87	135C 0020 0002	SINHALA LETTER AEYANNA
ඈ	0D88	135F 0020 0002	SINHALA LETTER AEEYANNA
ඉ	0D89	1360 0020 0002	SINHALA LETTER IYANNA
ඊ	0D8A	1363 0020 0002	SINHALA LETTER IYANNA
උ	0D8B	1366 0020 0002	SINHALA LETTER UYANNA
ඌ	0D8C	1369 0020 0002	SINHALA LETTER UUYANNA
ඍ	0D8D	136C 0020 0002	SINHALA LETTER IRUYANNA
ඎ	0D8E	136F 0020 0002	SINHALA LETTER IRUUYANNA
ඏ	0D8F	1370 0020 0002	SINHALA LETTER ILUYANNA

ඌ	0D90	1373 0020 0002	SINHALA LETTER ILUUYANNA
ඌ	0D91	1376 0020 0002	SINHALA LETTER EYANNA
ඌ	0D92	1379 0020 0002	SINHALA LETTER EYANNA
ඌ	0D93	137C 0020 0002	SINHALA LETTER AIYANNA
ඌ	0D94	1380 0020 0002	SINHALA LETTER OYANNA
ඌ	0D95	1383 0020 0002	SINHALA LETTER OYANNA
ඌ	0D96	1386 0020 0002	SINHALA LETTER AUYANNA
←Various Signs →			
◦	0D82	1389 0020 0002	SINHALA SIGN ANUSVARAYA
∴	0D83	138C 0020 0002	SINHALA SIGN VISARGAYA
← Consonants →			
ක	0D9A	1390 0020 0002	SINHALA LETTER ALPAPRAANAKAYANNA
ක	0D9B	1393 0020 0002	SINHALA LETTER MAHAAPRAANA KAYANNA
ග	0D9C	1396 0020 0002	SINHALA LETTER ALPAPRAANA GAYANNA
ස	0D9D	1399 0020 0002	SINHALA LETTER MAHAAPRAANA GAYANNA
ක	0D9E	139A 0020 0002	SINHALA LETTER KANTAJA NAASIKYAYA
භ	0D9F	139C 0020 0002	SINHALA LETTER SANYAKA GAYANNA
ච	0DA0	13A0 0020 0002	SINHALA LETTER ALPAPRAANA CAYANNA
ජ	0DA1	13A3 0020 0002	SINHALA LETTER MAHAAPRAANA CAYANNA
ජ	0DA2	13A6 0020 0002	SINHALA LETTER MAHAAPRAANA JAYANNA
ක	0DA3	13A9 0020 0002	SINHALA LETTER MAHAAPRAANA JAYANNA
ක	0DA4	13AC 0020 0002	SINHALA LETTER TAALUJA NAASIKYAYA
ක	0DA5	13AF 0020 0002	SINHALA LETTER TAALUJA SANYOOGA NAAKSIKYAYA
ජ	0DA6	13B0 0020 0002	SINHALA LETTER SANYAKA JAYANNA
ච	0DA7	13B3 0020 0002	SINHALA LETTER ALPAPRAANA TTAYANNA
ධ	0DA8	13B6 0020 0002	SINHALA LETTER MAHAAPRAANA TTAYANNA
ධ	0DA9	13B9 0020 0002	SINHALA LETTER ALPAPRAANA DDAYANNA
ධ	0DAA	13C0 0020 0002	SINHALA LETTER MAHAAPRAAN DDAYANNA
ක	0DAB	13C3 0020 0002	SINHALA LETTER MUURDHAJA NAYANNA
ධ	0DAC	13C6 0020 0002	SINHALA LETTER SANYAKA DDAYANNA
ත	0DAD	13C9 0020 0002	SINHALA LETTER ALPAPRAANA TAYANNA
ච	0DAE	13CA 0020 0002	SINHALA LETTER MAHAAPRAANA TAYANNA
ද	0DAF	13CC 0020 0002	SINHALA LETTER ALPAPRAANA DAYANNA
ධ	0DB0	13D0 0020 0002	SINHALA LETTER MAHAAPRAANA DAYANNA
න	0DB1	13D3 0020 0002	SINHALA LETTER DANTAJA NAYANNA
ද	0DB3	13D6 0020 0002	SINHALA LETTER SANYAKA DAYANNA
ප	0DB4	13D9 0020 0002	SINHALA LETTER ALPAPRAANA PAYANNA
ඵ	0DB5	13DC 0020 0002	SINHALA LETTER MAHAAPRAANA PAYANNA

බ	0DB6	13DF 0020 0002	SINHALA LETTER ALPAPRAANA BAYANNA
භ	0DB7	13E0 0020 0002	SINHALA LETTER MAHAAPRAANA BAYANNA
ම	0DB8	13E3 0020 0002	SINHALA LETTER MAYANNA
ඹ	0DB9	13E6 0020 0002	SINHALA LETTER AMBA BAYANNA
ය	0DBA	13E9 0020 0002	SINHALA LETTER YAYANNA
ර	0DBB	13EA 0020 0002	SINHALA LETTER RAYANNA
ල	0DBD	13EC 0020 0002	SINHALA LETTER DANTAJA LAYANNA
ව	0DC0	13EF 0020 0002	SINHALA LETTER VAYANNA
ශ	0DC1	13F0 0020 0002	SINHALA LETTER TAALUJA SAYANNA
ෂ	0DC2	13F3 0020 0002	SINHALA LETTER MUURDHAJA SAYANNA
ස	0DC3	13F6 0020 0002	SINHALA LETTER DANTAJA SAYANNA
හ	0DC4	13F9 0020 0002	SINHALA LETTER HAYANNA
ඌ	0DC5	13FA 0020 0002	SINHALA LETTER MUURDHAJA LAYANNA
ඔ	0DC6	13FC 0020 0002	SINHALA LETTER FAYANNA
← Dependent Vowels →			
ඞ	0DCF	13FF 0020 0002	SINHALA VOWEL SIGN AELA-PILLA
ඟ	0DD0	1400 0020 0002	SINHALA VOWEL SIGN KETTI AEDAPILLA
ච	0DD1	1403 0020 0002	SINHALA VOWEL SIGN DIGA AEDAPILLA
ඳ	0DD2	1406 0020 0002	SINHALA VOWEL SIGN KETTI ISPILLA
ඵ	0DD3	1409 0020 0002	SINHALA VOWEL SIGN DIGA IS-PILLA
ඹ	0DD4	140A 0020 0002	SINHALA VOWEL SIGN KETTI PAAPILLA
ඳ	0DD6	140C 0020 0002	SINHALA VOWEL SIGN DIGA PAAPILLA
ඞ	0DD8	140F 0020 0002	SINHALA VOWEL SIGN GAETTAPILLA
ඞඞ	0DF2	1410 0020 0002	SINHALA VOWEL SIGN DIGA GAETTA-PILLA
ඞ ඞ	0DD8 0DD8	1410 0020 0002	SINHALA VOWEL SIGN DIGA GAETTA-PILLA
ඟ	0DDF	1413 0020 0002	SINHALA VOWEL SIGN GAYANUKITTA
ඟ	0DF3	1416 0020 0002	SINHALA VOWEL SIGN DIGA GAYANUKITTA
ඹ	0DD9	1419 0020 0002	SINHALA VOWEL SIGN KOMBUVA
ඹ	0DDA	141A 0020 0002	SINHALA VOWEL SIGN DIGA KOMBUVA
ඹ	0DD9 0DCA	141A 0020 0002	SINHALA VOWEL SIGN DIGA KOMBUVA
ඹඹ	0DDB	141C 0020 0002	SINHALA VOWEL SIGN KOMBU DEKA
ඹ ඹ	0DD9 0DD9	141C 0020 0002	SINHALA VOWEL SIGN KOMBU DEKA
ඹඞ	0DDC	141F 0020 0002	SINHALA VOWEL SIGN KOMBUVA HAA AELA-PILLA
ඹ ඞ	0DD9 0DCF	141F 0020 0002	SINHALA VOWEL SIGN KOMBUVA HAA AELA-PILLA
ඹඵ	0DDD	1420 0020 0002	SINHALA VOWEL SIGN KOMBUVA HAA DIGA AELA-PILLA
ඹඞ	0DDC DCA	1420 0020 0002	SINHALA VOWEL SIGN KOMBUVA HAA DIGA AELA-PILLA
ඹ ඞ ඵ	0DD9 0DCF DCA	1420 0020 0002	SINHALA VOWEL SIGN KOMBUVA HAA DIGA AELA-PILLA

ඉ	0DDE	1423 0020 0002	SINHALA VOWEL SIGN KOMBUVA HA GAYANUKITTA
ඉ ො	0DD9 0DDF	1423 0020 0002	SINHALA VOWEL SIGN KOMBUVA HA GAYANUKITTA
ඊ	0DCA	1426 0020 0002	SINHALA SIGN AL-LAKUNA

8.2.3. Results

The collation elements were applied to sort a random set of strings of Sinhala. The input and corresponding output is given in Table 8.7.

Table 8.7. Input and Corresponding Sorted Output for Sinhala

Input		Output	
ඉරුව	කිකිටු	අංශ	කා
ඇළය	අක	අක	කාර
කටුව	ක	අකක්	කාරක
අංශ	කංසක	අකණය	කාරකය
කරස	කංසා	ආපසුපිකවා	කාරකයා
කටුවා	කංසාරිකි	ආපසුයකවා	කැ
උක්කණ	කංසක	ආපසුයාම	කැරළිකාර
ආපසුයකවා	කකණටකයා	ඇළය	කැරළිකාරක
ඇඳපු	කකළ	ඇඳපු	කැරළිභසකවා
කටු	කකා	ඉරුව	කෑ
කාරකය	කක්වයි	ඊගස	කි
කරවක	කවබැඳුම	උක්කණ	කිකමිද
කල්කණඩු	කවාරම	ක	කිකි
කාර	කවිකය	කංසක	කිකිටු
ආපසුයාම	කවු	කංසා	කිකුටු
කුබුදිකවා	කවෙ	කංසාරිකි	කී
කටුක	කා	කංසක	කූ
කරාබු	කැ	කකණටකයා	කුබුදිකවා
කෘමිල	කෑ	කකළ	කුබුදු
අකක්	කි	කකා	කුබුද්දිකවා
කොණ්ඩ	කා	කක්වයි	කූ
කැරළිකාරක	කො	කටු	කා
කැරළිභසකවා	කොඉ	කටුක	කාග
කිකමිද	කේ	කටුව	කෘමිල
කෙක්ක	කෘaa	කටුවා	කෘaa
කිකුටු	කඉ	කරරස	කඉ
කල්කියාව	කී	කරවක	කඉ
කුබුද්දිකවා	කූ	කරස	කෙ
කිකි	කූ	කරාබු	කෙකි
කාරකයා	කඉ	කල්කණඩු	කෙක්ක

කාරක කෘග කෙකි කැරලිකාර ආපසුපිකවා ඊගස	කෙ කේ කෙක අකණය කරඳස කුචුදු	කල්කියාව කවබැඳුම කවාරම කවිකය කවු කවෙ	කේ කෙක කො කොණ්ඨ කො කි
---	---	---	--------------------------------------

8.3. Conclusion

Sinhala has single level of collation, like other Indic languages. All characters are sorted at primary level. The sorting process requires some text processing to decompose the characters and map multiple characters onto single collation elements. However after the mapping, the collation algorithm discussed in the second chapter is applied in a regular manner for eventual collation.