PAN Localization

A Study on Collation of Languages from Developing Asia

Sarmad Hussain
Nadir Durrani

Center for Research in Urdu Language Processing
National University of Computer and Emerging Sciences

www.nu.edu.pk

IDRC CRDI
Canada
www.idrc.ca

Published by

Center for Research in Urdu Language Processing
National University of Computer and Emerging Sciences
Lahore, Pakistan

Printed by Walayatsons, Pakistan

ISBN: 978-969-8961-03-9

This work was carried out with the aid of a grant from the International Development Research Centre (IDRC), Ottawa, Canada, administered through the Centre for Research in Urdu Language Processing (CRULP), National University of Computer and Emerging Sciences (NUCES), Pakistan.

Preface

Defining collation, or what is normally termed as alphabetical order or less frequently as lexicographic order, is one of the first few requirements for enabling computing in any language, second only to encoding, keyboard and fonts. It is because of this critical dependence of computing on collation that its definition is included within the locale of a language. Collation of all written languages are defined in their dictionaries, developed over centuries, and are thus very representative of cultural tradition. However, though it is well understood in these cultures, it is not always thoroughly documented or well understood in the context of existing character encodings, especially the Unicode.

Collation is a complex phenomenon, dependent on three factors: script, language and encoding. These factors interact in a complicated fashion to uniquely define the collation sequence for each language. This volume aims to address the complex algorithms needed for sorting out the words in sequence for a subset of the languages. A small but diverse set of scripts and languages are chosen for this purpose from developing Asian region. The set is chosen for the variety it exhibits and to show the challenges it poses to solve the collation puzzle. Further details are given in the following chapters.

The data on different languages has been obtained from the dictionaries published in these languages, and through interacting with the PAN Localization project teams in relevant countries. First, the collation weights being proposed were developed based on analysis of the dictionaries and were implemented to sort out the words in these languages. Then results were verified with the dictionaries. Finally, the chapters written were reviewed by the relevant team members within the project. Thus, the results are both tested and verified. However, there is still more which can be said about collation of these languages. And, of course, there are many more languages which need to be documented. Accordingly, this work must be taken as an initial step towards addressing the collation of languages in the region.

We would explicitly want to acknowledge the support provided by all relevant PAN Localization country team members and their country project leaders for information and verification of the work. We would also like to thank International Development Research Center for the PAN Localization project grant, which has supported this work. Finally, we would like to thank the management of National University of Computer and Emerging Sciences for their continued support in research on language computing.

Sarmad Hussain
Nadir Durrani
(Lahore, Pakistan)

PAN Localization Project

Enabling local language computing is essential for access and generation of information, and also urgently required for development of Asian countries. PAN Localization project is a regional initiative to develop local language computing capacity in Asia. It is a partnership, sampling eight countries from South and South-East Asia, to research into the challenges and solutions for local language computing development. One of the basic principles of the project is to develop and enhance capacity of local institutions and resources to develop their own language solutions.

The PAN Localization Project has three broad objectives:

- To raise sustainable human resource capacity in the Asian region for R&D in local language computing
- To develop local language computing support for Asian languages
- To advance policy for local language content creation and access across Asia for development

Human resource development is being addressed through national and regional trainings and through a regional support network being established. The trainings are both short and long term to address the needs of relevant Asian community. In partner countries, resource and organizational development is also carried out by their involvement in development of local language computing solutions. This also caters to the second objective. The research being carried out by the partner countries is strategically located at different research entry points along the technology spectrum, with each country conducting research that is critical in terms of the applications that need to be delivered to the country's user market. Moreover, PAN Localization project is playing an active role in raising awareness of the potential of local language computing for the development of Asian population. This will help focus the required attention and urgency to this important aspect of ICTs, and create the appropriate policy framework for its sustainable growth across Asia.

The scope of the PAN Localization project encompasses language computing in a broader sense, including linguistic standardization, computing applications, development platforms, content publishing and access, effective marketing and dissemination strategies and intellectual property rights issues. As the PAN Localization project researches into problems and solutions for local language computing across Asia, it is designed to sample the cultural and linguistic diversity in the whole region. The project also builds an Asian network of researchers to share learning and knowledge and publishes research outputs, including a comprehensive review at the end of the project, documenting effective processes, results and recommendations.

Countries (and languages) directly involved in the project include Afghanistan (Pashto and Dari), Bangladesh (Bangla), Bhutan (Dzongkha), Cambodia (Khmer), Laos (Lao), Nepal (Nepali), Sri Lanka (Sinhala and Tamil) and Pakistan, which is the regional secretariat. The project started in January 2004 and has continued for three years, supporting a team of seventy five resources across these eight countries to research and develop local language computing solutions. The project is now entering its second phase, aimed to deploy the technology being developed, focusing on end user training and local language content, in addition to the language technology, for a wider set of languages across developing Asia. Further details of the project, its partner organizations, activities and outputs are available from its website, www.PANL10n.net.

Table of Contents

1. Introduction

We use lists on regular basis in our daily lives, such as shopping lists, address books and dictionaries. These lists are also used frequently by organizations and governments for employee payroll, bank accounts, telephone bills and voter lists. Most of these lists are ordered, i.e. their contents, which may include numbers, words or names, follow a specific sequence. The order depends on the cultural conventions, and is largely determined by two factors, the language in which the list is prepared and the script being used to write the language. This sequence is normally based on the order these words would occur in a dictionary or a lexicon and therefore, it is referred to as lexicographic order. It is also referred to as collation sequence.

The process of taking a list of randomly arranged words (or strings, as these are referred to in computing literature) and putting them in the collation or lexicographic sequence using a computer is called sorting. For example, in English, the words (or strings[1]) “mango”, “banana”, “apple” and “orange” are collated in the order ”apple”, ”banana”, ”mango” and ”orange” and the process which takes the initial set and arranges them in the latter sequence is called sorting. It is important to note that the collation sequence is linguistically and culturally determined, whereas sorting is a computational process to attain it. Thus, sorting is dependent on collation and not vice versa.

However, collation and sorting are not uniform across languages and scripts because of multiple reasons. First, various languages use different cultural conventions, even when using the same script. For example, ‘CH’ is taken as a single character between ‘C’ and ‘D’ in Spanish and ‘DZ’ and ‘DZS’ are taken as single characters between ‘D’ and ‘E’ in Hungarian [1]. Second, as different scripts use different orthographic units, e.g. phones, consonants, syllables, words, etc., the nature of collation can vary significantly as well. Chinese uses ideographs, Latin uses letters and Arabic uses consonantal sequences to represent words and thus employ different strategies in arranging words. Third, scripts may employ additional non-character elements to influence collation like capitalization, diacritics, vowel marks, tone marks, etc. For example, Arabic script uses vowel marks Fatha, Damma and Kasra which (similar to capitalization in Latin script) influence order only if base character string is same. Thus, for Urdu بَن precedes بِن which precedes بن. Fourth, there are always exceptions to the rules, which are culturally fossilized, for reasons which may or may not be known. Finally, novel ways of generating newer strings are being introduced, which cross beyond the conventional collation boundaries and thus existing collation rules have to be extended. For example, R2D2 and C3PO are names of two robots

[1] Strings are made of zero of more characters. Characters are normally enclosed within single quotes and strings are enclosed within double quotes. In this document, the quotes are only put in where there is a chance of ambiguity.

introduced in Star Wars which are alpha-numeric strings, though normal collation only deals with alphabetic strings[2]. Interestingly, a language may employ multiple collations and consequently sorting methods as well. For example, Chinese can be sorted based on Latin transcription[3], phonetics[4], character shape or radical and stroke count [2]. In such cases, all the expected collations must be realized and the choice of a particular sort must be left on the user. A detailed discussion on these aspects is given later in this report.

This report looks at a variety of languages and writing systems used in developing Asia to bring out the diversity and the challenge associated with collation in this region. The study discusses the following languages (and writing systems): Bengali (Bengali), Dzongkha (Tibetan), Lao (Lao), Mongolian (Cyrillic), Sindhi (Arabic), Sinhala (Sinhala), Tamil (Tamil) and Urdu (Arabic). Bengali writing system is closely related to Devanagari writing system. Sinhala and Tamil writing systems derive from Brahmi script, as used for South Asian languages (Indo-Aryan and Dravidian respectively). Lao also derives from Brahmi script (but related to Khmer and Thai), and represents a tonal South-East Asian language. Sindhi and Urdu are both based on different extensions of basic Arabic script.

The report first introduces the linguistic and technical aspects relevant for collation and presents the Unicode Collation Algorithm as one of the sorting techniques for multilingual strings. Then the report individually discusses the collation of languages and their peculiarities, presenting the solutions for their collation. The report concludes with a comparative discussion on collation of the languages discussed, also identifying where more work needs to be undertaken.

[2] A more contemporary chatting domain uses words like “b4” for “before”, introduces new spellings like “u” for “you” and collapses multiple words into single words e.g. “lol” for “laughing out loud” and “gtg” for “got to go”.

[3] Also known as Pinyin system [3].

[4] Also known as Bopomofo. It is used in Taiwan [1].

2. Collation

2.1. *Encoding and Collation*

Computers inherently process numbers. However, we frequently require them to manipulate words of natural languages. To represent these languages within the computer, all the characters in each of the languages are associated with a unique number. American Standard Code for Information Interchange (ASCII) was first introduced as a representation scheme for English characters. In ASCII ‘A’, ‘B’, …, ‘Z’ are assigned numbers 65, 66, …, 90. Similarly, ‘a’, ‘b’, …, ‘z’ are assigned numbers 97, 98, …, 122. ASCII is a mono-lingual standard and can only fit in 128 characters (seven bits). This mapping of 128 characters to numbers or codes is normally also referred to as a code page. English code page was eventually extended to include other languages (each code page including 256 characters or eight bits), e.g. the ISO 8859 standard. With the advent of multilingual computing, where a single piece of text may contain multiple languages, a new text encoding standard was developed, which has initial space 27 bits and is called Unicode or ISO IEC 10646 standard [4]. This is a script based standard and encodes each character of each script uniquely[1]. Unicode is currently the most widely supported and commonly used multilingual standard.

When the strings are sorted by the computer, one choice is to base it on character codes. This would sort “cat” before “dog”. However, as is apparent from ASCII codes given above, it would sort “Zebra“ before “cat“ and “dog“, which is incorrect according to the English lexicographic order. What if the characters are encoded such that they come in the collation order: ‘A’, ‘a’, ‘B’, ‘b’, ‘C’, ‘c’, …, ‘Z’, ‘z’. Could then the codes used for encoding be used for sorting? No, because it would still not tackle the case and sort ‘Mango’ before ‘man’. Moreover, same codes are also used for other languages using Latin script, for which it would not work, e.g. for letters ‘CH’ for Spanish and ‘DZ’ and ‘DZS’ for Hungarian as discussed. Consequently, a single script level encoding cannot be used to properly collate all languages which use the script. Same is true for other scripts. For example from Arabic script, Urdu requires the sequence (from right to left): ب پ ت ث, while Sindhi requires the sequence (from right to left): ب ت ث پ [5]. And from Devanagari script, Hindi requires the sequence ल, ळ, व whereas Marathi requires the sequence: ल, व, ळ [1].

This illustrates a very fundamental principle that character encoding and collation are independent phenomena and therefore character codes cannot be used for sorting. Hence, a separate set of codes is normally assigned for sorting characters. These codes, different from

[1] If two languages use the same character, they will use the same Unicode, as this standard does not encode on the basis of language but on the basis of script

character encoding, are called collation elements or collation codes. Though character encoding is done for scripts, the collation elements must be defined separately for each language because collation is a language specific phenomenon. English, French and Spanish may share the same character encoding, but will each have its own set of collation elements. Similar would be the case of Nepali and Hindi, even though they are both written in Devanagari script. Using these language specific codes, it is possible to sort the strings in correct lexicographic order of the language. Moreover, for languages which have multiple collations, like Chinese, multiple sets of collation elements will be required; one set for each collation strategy.

Even though character encoding is logically independent of collation elements, the two are still intricately connected. For example, though ‘CH’ is a single character in Spanish, it is not differently encoded in Latin encoding in Unicode, it does not exist and must be defined as a combination of ‘C’ and ‘H’ and must be processed accordingly. Similarly, ب and ﻫ are separate characters in Arabic language but combine to make a single character به in Urdu language.

The next sections discuss details of these linguistic, orthographic and encoding related complexities related to collation as observed for various languages, highlighting their interconnection and practical dependencies.

2.2. Starting with Collation

Work on collation starts with linguistic and orthographic analysis of a language. In this analysis, the first step is to determine the complete character set of the language, which includes letters, digits, punctuation marks, arithmetic marks, other marks, other letters, etc. (e.g. see [8, 9] for Urdu language; similar references to work on other languages are given in [3]). Second, the sub-set of this character set which plays a role in collation needs to be identified and separated from characters that do not participate in collation. The third step is to determine which factors influence collation for these characters and how these factors interact with each other to sort the characters. For example, Wissink and Kaplan [1] list casing, modifier marks, syllable structure, pronunciation and stroke-count as some of these factors which influence word order in various languages. These and additional factors are discussed in more details in Section 2.3.

After the linguistic and orthographic details are understood and documented the technical details also need to be analyzed. First, the character set has to be completely encoded. If existing encoding (e.g. Unicode) is to be used, any missing characters must be added to the encoding. Once the character set and its encoding has been verified, the text processing required for collation needs to be determined and developed. This may include normalization, ordering and reordering, contraction, syllabification and other related processes, which are discussed in more detail in Section 2.4.

Finally, after pre-requisite analysis has been conducted and details are finalized, collation elements for the language have to be defined to enable sorting. Details of this process and associated algorithm are discussed in the Section 2.5.

2.3. Linguistic and Orthographic Factors

Linguistic and orthographic factors are those which are inherent in the behavior of the language and its writing system, independently of the encoding scheme. The following sub-sections explain some of these factors that are directly relevant for collation. Different languages and scripts employ a combination of these factors. However, as illustrated with examples, the way these factors are used varies across languages and scripts.

2.3.1. Collation Levels

The sub-set of characters which are relevant for collation from the complete character set has to be identified. Similarly, the ignorable characters, which are not relevant for collation, also need to be identified. For example, letters ‘A’, ‘B’, …, ‘Z’ and ‘a’, ‘b’, …, ‘z’ are all relevant for collation. Characters ‘(‘, ‘)’, ‘[‘, ‘]’, ‘$’, ‘+’ are all part of English character set but are ignorable for collation purposes. For Urdu language, using Arabic script, letters آ، ب، پ، ت، --- ے are all relevant characters for collation. The marks َ، ِ، ُ، ّ are also relevant. However, punctuation marks ؛، ؟، -، : are irrelevant.

Furthermore, the linguistic and orthographic phenomena applicable in the language need to be identified and prioritized. For example, English collates on letters, i.e. “apple” sorts before “banana”. Where characters are the same, it further sorts on casing, i.e. “Apple” sorts before “apple”[2]. Thus, A < a < B < b for English. French is different from English because it sorts first on letters, then on marks and finally on casing[3]. Thus, the sequence for French characters is as follows: A < a < Á < á < B < b. French is further exceptional because the marks are collated from right to left (from the end of the word towards the beginning). This is further different from Swedish where marks define a different character, but casing is still treated at the secondary level. Thus, A < a < O < o < Y < Z < Å < Ä < Ö is followed for Swedish [10]. In Urdu, which uses Arabic script, letters carry the primary significance, followed by a subset of marks which represent

[2] Assuming ‘A’ sorts before ‘a’.

[3] Assuming that English does not have marks, and words like “naïve” are exceptions.

consonantal and vocalic segments. Urdu also uses a second class of marks, called honorifics [9] which do not add any consonantal or vocalic material but add respect to the words[4]. These marks add the tertiary significance to collation. Thus, علی would come before □ علی, latter having an additional honorific mark on the last letter. Lao has four levels of sorting. Primary sorting is based on central consonants. The other levels are based on vowels, consonantal marks and tone marks respectively. Lao further uses a syllable to do sorting properly. Depending on the place in a syllable, the same character may represent a different level of collation. Thus, the significance of a character in sorting is context dependent. Unicode collation allows for multiple levels of sorting. For each language, it needs to be identified how many levels are required and which property is relevant at these levels. The factors which determine these levels are discussed in the next sections.

2.3.2. Casing

Many scripts, including Latin, allow letters to take upper and lower cases. Thus, in English the string “apple” and “Apple” mean the same but are orthographically different. “Apple” is placed before “apple” in lexicographic order, but both will occur before “Banana” and “banana”. This means that whether in upper case or lower case, the letter ‘a’ (or ‘A’) precedes the second letter ‘b’ (or ‘B’). However, within the same letter, upper case letter comparison ‘A’ precedes lower case ‘a’. Thus, casing influences the collation order, but not at the same level as a character. No matter what the case, first the character order A, B, C, D, …, Z for English is observed. Where characters are the same, the case is further used to sort the words. Thus, in Latin script based languages, casing influences sorting but not at the primary (character) level. Casing is also used in Cyrillic and Greek scripts.

2.3.3. Marks

Most languages add marks to base characters to add further information. Collation is also sensitive to this additional information. However, nature of marks differs greatly across scripts and languages and thus their implication on collation also varies. Most of the times these marks are used to represent or modify pronunciation, and represent vocalic and consonantal features (e.g. in Latin, Arabic and Indic). In other cases, these marks also represent supra-segmental features like tone (e.g. in Lao and Vietnamese).

Starting with Latin script, though English language does not take on any marks, many European languages using this script use marks abundantly. Some of these languages include French, Swedish, Danish, Turkish and Finnish [1, 10]. Marks are also abundantly used in Asian

[4] Mostly used in religious context by Muslims.

languages using Latin script, e.g. Vietnamese and Malay. However, each language may use these marks differently, for example Vietnamese uses marks to represent tone [11]. Some of these Latin marks include acute accent, grave accent, diaeresis and circumflex as shown on capital ‘A’ respectively: À Á Ä Â. These marks influence the sorting in different ways. In most of these languages, the characters still carry the primary importance for sorting, followed by the marks. Casing is less important than marks and thus carries a tertiary level significance. French language is unique as it sorts the words with characters as they appear from left to right, but accents as they appear from right to left [1, 5]. However, not all languages using Latin script treat marks as secondary. As reported in [10], Swedish (A < B < Y < Z < Å < Ä < Ö) and Danish (A < B < Y < Z < Ä < Ö) treat characters with marks as different characters and sort them at primary character level.

Marks are also used to represent vowels in Arabic script based languages. These marks play a secondary role in collation. Thus, in Urdu language, the letters are sorted first with Fatha, then Kasra and finally with Damma, as shown for letter بَ بِ بُ:ب (read from right to left). There are additional marks as well, which influence collation. There are also additional marks used in Arabic script based languages, which have no influence on collation and are ignorable in this context.

Indic languages also use marks. For example, in Devanagari script used to write South Asian languages like Hindi, Nepali and Marathi, Chandrabindu, Anusvara and Visarga when used with a base character, make the combination sort before the base character in the given order, as shown for the letter क: कँ कं कः क. However, Nukta mark behaves differnetly and sorts the combination after the base character without a mark, giving the following order for the same letter: क क़ [10]. Bangla, Tamil, Sinhala and other scripts used in South Asia show similar behavior.

Marks are used by other scripts as well. For example, South East Asian scripts Lao, Khmer, Thai and Burmese use marks to represent vowels, tones and/or other linguistic phenomenon, which play a role in their collation as well.

In summary, marks are used in orthography by most scripts to represent a variety of linguistic and other phenomena. These marks influence collation in most cases (though do not influence collation in some cases). However, the influence may be at primary, secondary or tertiary level. This level of influence is not consistent across scripts. Some languages may use marks at primary level, while other may use them at secondary level even if these languages use the same script. In addition, the level of influence may also vary within a language. A particular mark may be used at a secondary level while another mark may be used at a tertiary level within a language. Each language must be investigated to determine the role of marks.

2.3.4. Syllables

Though most textual analysis in many languages is based on characters as they are organized in words, some languages also use units larger than characters but smaller than words for internal structuring. In most cases these units align with syllables[5], though in other cases, the mapping is motivated by syllables (which are phonological and based on sounds) but is based on orthography, creating clusters of letters (not sounds).

In Lao language, the text comparison is not done at character level. The string is divided into syllables and then the sorting is done on syllable-wise comparisons. There are no explicit syllable markings and syllabification needs to be computationally done through a complex set of rules (details discussed in the chapter on Lao language later). Dzongkha, on the other hand, uses syllable level analysis but marks the syllable boundaries explicitly using Tsheg mark ་ (U+ 0F0B). Syllabification is also required for Urdu using Arabic script, but this syllabification is orthographically motivated for text processing and does not align with phonological syllables. Thus, the word بَن is a single phonological syllable (CVC) but may be sub-divided into two orthographic entities بَ and ن (CV and C respectively) in some cases for text processing. This sub-division of words into phonological or orthographic constituents larger than letters is relevant to collation for some languages.

2.3.5. Other Linguistic and Orthographic Factors

Various scripts and languages also employ additional factors used for text processing, including collation. Chinese uses different ways of sorting, based on pronunciation of words (called Bopomofo or Zhuyin fuhao) or the number strokes used to write the words (called stroke count) [1]. Chinese may also be sorted on encoding order, e.g. based on Unicode or BIG5 encoding [3]. Arabic normally uses two or three consonants for a morpheme which is realized as different words through vocalic infixation. For example, sequence of ک ت ب (K T B) represents the morpheme “book” and various words on this concept are realized by infixing different vowel sequences, e.g. کِتاب (KiTaB, “book”) and کتب (KuTB, “books”). Changing the vocalic infix makes

[5] Syllable is a linguistic entity which is well defined in Phonology, e.g. see [7].

inflectional or derivational changes in the word[6]. In some cases, the Arabic words may be sorted based on the underlying consonantal template irrespective of the actual surface word.

Not all languages have a complete algorithmic way of collating all words, and may also be arbitrary for some words based on traditional use of the language. For example, some languages use traditional dictionaries developed hundreds of years ago as the reference to organize words, e.g. Choun Nat dictionary for Khmer language in Cambodia. Though such dictionaries do show patterns in arrangement of words, it may not be true for all words in these dictionaries and some arbitrary ordering may be necessary to meet cultural expectations.

Thus, a variety of linguistic and orthographic phenomena interplay to define collation and other related behavior for a language. This is language specific and is not consistent across script or geographical regions. There may also be multiple ways of collating strings within the same language.

2.4. Text Processing

Normally the raw input string of a language undergoes initial processing before it can be sorted. Though it is fundamentally based on linguistic and orthographic characteristics, it is also critically dependent on the way the language is encoded. This section discusses encoding related phenomena for text processing. All these processes are not applicable to all languages.

Some additional processes, not directly applicable to collation, are also explained. This has been done because sometimes they are confused as having implications on collation and thus clarification on the disconnection is required. As Unicode is the default standard for multilingual encoding, all discussions in this and later sections is based on this standard (see [4] for details on Unicode[7]). However, similar pre-processing may also be needed for other encodings. Once the string is processed, it is then assigned the collation elements and actual sorting is performed. This second step in discussed in the next section.

2.4.1. Text Input and Rendering

Once encoded, multiple methods may be employed to input the text from the user for a language. These include simple typing using a keyboard for English to much more complex handwriting

[6] Also known as templatic morphology. Morpheme is the underlying form of a word, latter being a surface form. Infixation inserts letters inside the morpheme rather than before or after it, as in the case of prefixation or suffixation respectively.

[7] Details are also available at www.unicode.org.

recognition based systems for Chinese. Collectively, these are called *input methods* and take user input in form of keystrokes, speech or hand-writing and convert it into a series of letter codes based on the encoding (e.g. Unicode), which are eventually stored internally for further processing. See [3] for further details and references.

Once the text is input into a computer, it may also be displayed on the screen or *rendered* for users to view. This is done through a software program called the rendering engine, which uses the input and associated font files to generate the visual output on the screen. Same encoding but a different font file can cause cosmetic changes in the way output looks (e.g. Times New Roman vs. Courier New fonts). It is important to note that in the rendering process, choice of font and the output does not change the internal encoding of the text.

Some writing systems are also context-sensitive. Thus, the same letter may have a different shape depending on where it occurs. For example, in Arabic script, a letter takes a different shape if it occurs in initial, medial, final or isolated position in a connected portion[8] of text. Thus, the string ششش represents initial, medial and final shapes of the same letter ش in the connected portion (the text should be read from right to left). These different shapes are realized through the font and rendering system but represent the same underlying code.

For collation and related text processing, input methods, rendering and context sensitive shaping are not relevant[9]. This processing only depends on encoding.

2.4.2. Text Direction

Scripts use different writing directions. Latin, Greek, Devanagari and many more are written from left to right. Arabic and Hebrew are written from right to left. Similarly, Mongolian and Chinese are written from top to bottom. Sometimes two text directions may also be mixed. For example, in Arabic and Hebrew letters are written from right to left but digits are written from left to right, and are thus called bi-directional scripts.

Even though the text may appear in multiple directions on the screen, it is only stored in the key-press order internally, i.e. the order in which the individual characters are keyed in or written by the user. The visual order is not relevant as collation is done based on internal storage which

[8] Arabic writing is cursive and thus letters are joined together when written. Connected portion is also referred to as a ligature.

[9] Unicode has encoded some context sensitive shaping, e.g. for Arabic script, for backward compatibility. However, use of this area is not encouraged and not discussed here.

uses the key press order. Thus, the text direction does not have any implication on collation algorithms.

2.4.3. Normalization

Due to various reasons, e.g. compatibility with legacy encoding systems or re-use of productive combining marks, there may be multiple ways of representing the same character in Unicode. For example, the letter ë (U+00EF[10]) may also be represented by e (U+0065) followed by Diaeresis ¨ (U+0308). There can also be multiple possibilities. For example, the letter ǖ (U+01D6) may be represented by ü (U+00FC) followed by Macron ¯ (U+0304) and by u (U+0075) followed by Diaeresis ¨ (U+0308) and Macron ¯ (U+0304) [5]. The encoding is also redundant for other scripts. For example, in Devanagari script the letter ऱ (U+0931) is same as the sequence र (U+0930) in combination with the mark ◌़ (U+93C), and Arabic script letter آ ((U+0622) is same as the sequence ا (U+0627) in combination with the mark ٓ (U+0653). As can be seen, most redundancies result from the fact that combined characters, base characters and combining marks are all encoded within the standard.

This redundancy in encoding can cause problems in processing. For example, if a spell checker verifies on the code of composed form ऱ for a language, and some text uses the de-composed form र + ◌़, then the spell checker may give and error where there is none. Similarly, a search engine may not be able to find the Urdu word آم in a text corpus if it is searching for آ and ا + ٓ is encoded. Thus, the text has to be brought in a consistent format for the eventual processing needs. This conversion of text into consistent representation is called *normalization*. Normalization may be done either to totally compose the characters, where de-composed forms are possible, or alternatively totally decompose the characters, where composed forms are possible. Either can work effectively, as long as consistency is maintained.

Complete decomposition may require multiples steps, as each step takes off a single mark from the combination and sometimes a composed character may contain multiple marks, as has been seen for the case of the letter ǖ. Also, composition and decomposition may also require some additional re-ordering steps to get the individual elements within a composition to come in an expected or stipulated sequence. This is discussed in the next section. For details of normalization for Unicode see [27].

[10] Unicode consortium make use of 'U+x' notation to express Unicode code points where x is a 4-6 hexadecimal digits, using the digits 0-9 and upper case letters A-F (for 10-15 respectively) [4]

Once the text is normalized (into either composed or decomposed forms), it can be further processed for collation, as discussed later.

2.4.4. Ordering and Reordering

With Unicode encoding, the Vietnamese letter ộ (U+1ED9) may be represented in five ways [5]: (i) o +ˆ+ ̣ , (ii) o + ̣ +ˆ, (iii) ô + ̣ , (iv) ọ +ˆ and (v) ộ , where '+' represents concatenation of the symbols. This presents a problem for processing text encoded with Unicode, as different users may write the same string in many different ways. To address this concern, Unicode assigns a *combining class*, a number from 0 till 255, to each character which can be used to determine their ordering (done in increasing order). Low numbers are assigned to those characters or combining marks which come first. Non-combining characters are assigned 0. Among the combining characters, the marks which are placed below are arbitrarily assigned lower numbers compared to marks which are placed above base the characters. Order of characters or marks within the same combining class is not changed. So, for the above example, under-dot has a combining class number 220 and the circumflex above has the combining class number 230. Thus, the canonical decomposed form defined by Unicode would be (ii) above and not (i). Any processing of decomposed forms of a combination of characters must first take the input string, completely decompose its contents and adjust the *ordering* as described before further processing. Similar is true for other languages, e.g. Lao stacks vowels and tone marks above and some combining characters below the base glyph, e.g. ທີ່ and ຍ່◌ຼ, and would follow same recommendations for canonical order.

Many Indic writing systems are syllabic and are written in a consonant-vowel (CV) combination. Consonant always logically precedes the vowel, though some vowels appear before the consonants (even though they are typed after the consonant). These are called left-combining vowels in this left-to-right writing system. Unicode also encodes these languages to follow this logical CV typing order, except in Lao and Thai, where logically the language is processed in a similar fashion but encoding order is visually based, i.e. left-combining vowels are typed before the consonant rather than after it[11] as VC. Thus, for other processing like sorting, which expects logical order CV, the typing order has to be reversed before processing. This is normally referred to as *reordering*. Reordering is only done on a copy of the input string for further processing and does not alter the actual input (latter is required for proper rendering on the screen).

[11] This has been done to keep the encoding compatible with national standards and legacy encodings.

2.4.5. Contraction

Often two or more characters clump together to form linguistic unit which has its own identity in collation or other string manipulation processes. This group is treated similarly as a single character. These units may not be directly encoded in Unicode but are required to be created from their constituent units which are encoded. This process is called *contraction*. For example Spanish has a unique character ‘CH’ (U+0043 + U+0048) different from ‘C’ and ‘H’. It sorts between C and D [1] in Spanish. So C and H occurring together in an input sequence are required to be collapsed onto a single collation element. Contraction also occurs in Arabic script based languages, for example in Urdu the letter بھ is formed by contracting the letters ھ (U+06BE) and ب (U+0628). This combined letter is not separately encoded in Unicode.

.

2.4.6. Context Sensitivity

Though many languages, especially cursive languages, change the shape of the letter depending on its context as discussed earlier, some languages also change the ‘behavior’ of the letter. This change is not just cosmetic and changes the way the letter has to be processed. For example, in Spanish, if ‘C’ is followed by an ‘H’ they combine to form a single letter ‘CH’, but otherwise behave as individual letters. In Naskh writing style of Arabic script letter ھ may represent independent character Hay or be part of the previous consonant to represent the aspirated sounds (e.g. بھ، کھ، جھ etc.)[12]. Similarly in Lao the letter ນ could be the nuclear consonant in a syllable but may also be a dependant/alternate consonant if it follows another nuclear consonant, e.g. in the sequence ຫນ.

Dzongkha presents even a more complex scenario. In Dzongkha consonants very productively conjoin to form larger linguistic units, though only context can determine which of the conjoined letter heads the cluster. There are some letter clusters that depend on the third or fourth subsequent letter to decide whether it is root letter prefixed by another letter or a root letter followed by a suffix. For example, in the sequence དག it is hard to decide if ད is prefixed with root ག or ག is suffixed with root ད without looking at neighboring character which itself is not part of the cluster. This dependence of behavior on neighboring letters is referred to as *context sensitivity* and is computationally complex to model. However it is required as collation is based on root character.

[12] Normally Nastalique style of writing is used for Urdu, which does not have this ambiguity.

2.5. Collation Elements and Sorting

Once all characters, their linguistic and orthographic properties and encoding specific requirements are addressed, each character is assigned a weight, called collation element. The collation element is eventually used to sort the strings.

2.5.1. Collation Elements

A collation element is a weight assigned to a character which is used to compare it with other characters in the sorting process to determine which character is 'lighter' in a pair-wise comparison. This collation element is further divided into a set of numbers or weights, specifying significance at different levels. Unicode collation algorithm [2] uses four weights to define significance at primary, secondary, tertiary and quaternary levels, but may be extended if needed for a language. Sample collation elements are shown for some languages in Table 2.1 below.

Table 2.1. Collation Elements

Glyph	Unicode	Collation Elements	Unicode Name
R	0052	[09CB 0020 0008 0052]	LATIN CAPITAL LETTER R
Д	0414	[0E2D 0020 0002 0414]	CYRILLIC CAPITAL LETTER DE
д	0434	[0E2D 0020 0008 0434]	CYRILLIC SMALL LETTER DE'
ක	0D9A	[1390 0020 0002 0D9A]	SINHALA LETTER ALPAPRAANAKAYANNA
ກ	0E81	[0000 0000 0025 0E81]	LAO LETTER KO
◌ُ	064F	[0000 00CB 0002 064F]	ARABIC DAMMA
◌໋	0ECB	[0000 0000 0000 0011]	LAO TONE MAI CATAWA

Template of a collation element is [w1 w2 w3 w4], where w1, w2, w3, and w4 represent the weight of the character for collation at primary, secondary, tertiary and quaternary levels respectively in form of hexadecimal numbers[13]. For example, characters ‘A’ ‘B’ and ‘C’ in English will have a non-zero w1 to indicate that these characters participate in collation at primary level. Character ‘A’ should also have a smaller w1 than ‘B’ to indicate that A < B in English. The weight w1 for characters ‘A’ and ‘a’ should be same as they represent the same character at primary level. The difference in casing has a tertiary level effect. Thus, ‘A’ and ‘a’ will differ in w3, with ‘A’ having a smaller tertiary weight if A < a for English.

A weight of 0 indicates that the weight should not be considered. A character which does not participate in collation should have w1=w2=w3=w4=0. If a character participates at secondary level, as is the case for Arabic mark Damma in Table 1, its w1=0, which indicates that it should be ignored at primary level. Lao tone which collates at quaternary level has all weights equal to zero except w4. This makes the tone ignorable at first three levels. It is important to note that if a character is ignorable at a level, it must also be ignorable at all levels before it. Thus, if w2=0, then w1 must also be zero. It also implies that if a character is not ignorable at primary level, its other weights w2, w3, w4 must also be non-zero values.

It is not important what the exact value of w1, w2, w3, w4 should be except that they must reflect the comparative sequence within characters of a language at each level. Thus, in Urdu there are multiple secondary level marks, Fatha, Kasra and Damma, and are collated in this order. Each has w1=0. It is not important what the exact values of w2 are as long as w2(Fatha) < w2(Kasra) < w2(Damma). See [2] for further details and further explanation.

A language specific table for all characters needs to be defined in this manner for each language. The table should give the following details: (a) Character(s) which map onto a single collation element, (b) Unicode of the character(s) for which the collation element is being defined. Multiple characters may map onto the same collation element, if cases for contraction have been identified for the language (as discussed above), (c) Corresponding collation element with w1, w2, w3 and w4 specified to represent the level and order of collation for the character(s), (d) Unicode name(s) of the character(s), (e) Optionally, any explanatory notes. Columns (a), (b) and (d) should also list all the alternate possibilities of characters which may also map onto the same collation element (e.g. if Unicode encodes multiple ways of representing the same character, as discussed in section explaining Normalization, or in case multiple characters map onto a single collation element, as discussed in the section explaining Contraction). Examples from Urdu in Arabic script are given in Table 2.2 below.

Table 2.2. Collation Elements for Normalization and Contraction

[13] Each four bit hexadecimal number represents a 16-bit binary number.

Glyph	Unicode	Collation Elements	Unicode Name
ﻻ	FEFB	[13AB 0020 0002][1350 0020 0002]	ARABIC LIGATURE LAAM WITH ALEF ISOLATED FORM
آ	0622	[1350 0020 0002][0000 00F1 0002]	ARABIC LETTER ALEF WITH MADDAH ABOVE
بھ	0628 06BE	[1353 0020 0002]	ARABIC LETTER BEH + ARABIC LETTER HEH DOCHASHMEE

This is background research and development which has to be taken up for each language before actual strings can be sorted. Details of the sorting process are given below.

2.5.2. Default Collation Elements

Unicode standard aims to concurrently encode all written scripts. It is a multilingual standard which is targeting to encode the scripts in such a way that they may be inter-mixed. However, when inter-mixed text is collated, it may encounter characters which are not assigned any collation elements. This may produce unexpected results. To overcome such challenges, Unicode recommends using Default Unicode Collation Element Table (DUCET) [16] as a backup if collation element is not found in a language table, latter taking the precedence. Collation elements given in DUCET do not collate the characters for any particular language and thus do not give the right collation sequence. They just allow for a dependable and consistent wrong sorting.

2.5.3. Sorting Words

Once the collation has been set up for a language, the following algorithm is followed to sort two words or strings. Each step is applied to both candidate words for final collation. The same can be extended to larger list of words. This algorithm is followed by an example, both adapted from [5] and are based on the Unicode collation algorithm [2]. Even if other algorithms are used, they will fundamentally be similar to what is done in Unicode Collation Algorithm at least at linguistic and orthographic levels.

(i) Take individual characters in the input and determine if any of the characters need to be decomposed by consulting the decomposition table for the language.

(ii) Reorder any characters, in case reordering is required.

(iii) Assign collation elements for each character or sets of characters in the sequence they appear in the input. Multiple characters in a sequence can get a single collation element in case contraction is being done.

(iv) Group the character weights for the complete word by levels, ignoring zero weights and inserting an additional zero between each level to form a single sort key.

(v) Compare the sort key with the sort key of other word to determine the collation order.

Consider sorting two words, “Resume” and “résumé”. In Step (i), we take the following characters:

Resume:	0052 0065 0073 0075 006D 0065
résumé:	0072 00E9 0073 0075 006D

and decompose é in the second word to e +´, thus getting the Unicode sequences:

Resume:	0052 0073 0075 006D 0065
résumé:	0072 0065 0301 0073 0075 006D 0065 0301

No reordering of characters is required in English, so Step (ii) is not applied. Next, the collation elements corresponding to the Unicode codes are obtained from a mapping table developed for English. Each code is replaced by its collation element, as in Step (iii)[14].

Resume:	[09CB 0020 0008] [08B1 0020 0002] [09F3 0020 0002] [0A23 0020 0002] [0977 0020 0002] [08B1 0020 0002]
résumé:	[09CB 0020 0008] [08B1 0020 0002] [0000 0032 0002] [09F3 0020 0002] [0A23 0020 0002] [0977 0020 0002] [08B1 0020 0002] [0000 0032 0002]

After accessing the collation elements for each character within a word, the collation keys are formed by grouping the non-zero weights of each character in the word at a level, separating levels by an additional zero, as in Step (iv).

Resume:	09CB 08B1 09F3 0A23 0977 08B1 0000 0020 0020 0020 0020 0020 0020 0000 0008 0002 0002 0002 0002 0002
résumé:	09CB 08B1 09F3 0A23 0977 08B1 0000 0020 0020 0320 0020 0020

[14] Three levels of weights are given in this example. The same algorithm can be extended to any arbitrary number of levels.

```
0020  0020  0032 0000 0002 0002 0002 0002 0002 0002 0002 0002
```

The sort keys thus formed are compared to determine which word is sorted first. The comparison shows that the words are equal in value until the third value in the secondary weight caused by the accent on the first ‘e’ of “résumé”. This makes the second word “heavier” and sorts it after the word “Resume”[15]. In this case, the difference of ‘R’ vs. ‘r’ in the two words is not considered to make the decision as its effect comes later in the sort key at the tertiary level.

[15] It is being assumed that accents have secondary weight and casing has tertiary weight in English in this example.

3. Bengali

Bengali (ethnonym: Bangla) language is categorized within the Bengali-Assamese branch of Eastern Zone of Indo Aryan languages. It is spoken by more than 200 million people across the world out of which about 100 million speakers reside in Bangladesh and 70 million speakers reside in India [12]. Bengali is the national language of Bangladesh while it is also the state language of the Indian state of West Bengal.

Bengali is an Indic language which uses Bengali script, closely related to Devanagri script, both deriving from Brahmi script. Bengali script is also used to write other languages, including Assamese, Daphla, Garo, Hallam, Khasi, Manipuri, Mizo, Munda, Naga, Rian and Santali [4, 13].

3.1. Writing System

3.1.1. Character Set

Bengali character set is divided into 21 vowels, 36 consonants and modifiers [15]. The vowels themselves can be divided into dependent and independent vowels, shown in Figure 3.1 below.

অ আ ই ঈ উ ঊ ঋ এ ঐ ও ঔ

Independent Vowels

◌া ◌ি ◌ী ◌ু ◌ূ ◌ৃ ◌ে ◌ৈ ◌ো ◌ৌ

Dependent Vowels

Figure 3.1. Bengali Vowels

ক খ গ ঘ ঙ চ ছ জ ঝ ঞ ট ঠ ড ঢ ণ

ত থ দ ধ ন প ফ ব ভ ম য র ল শ ষ

স হ ড় ঢ় য় ৎ

Figure 3.2. Bengali Consonants

Along with consonants and vowels there are some special modifiers, called Virama, Visarga, Anusvara, Candrabindu and Ishar, shown in Table 3.1. Anusvara is used for final velar nasal sound, Visarga adds voiceless breath after vowel and Candrabindu is used to nasalize vowels [13, 14]. Virama, also called Halanta is discussed in the next section.

Table 3.1. Bengali Special Characters

Name	Glyph	Usage with a Consonant ‘k’
Virama	◌্	ক্
Candrabindu	◌ঁ	কঁ
Anusvara	◌ং	কং
Visarga	◌ঃ	কঃ

Bengali also has its own numerals, shown in Figure 3.3.

০ ১ ২ ৩ ৪ ৫ ৬ ৭ ৮ ৯

Figure 3.3. Bengali Numerals

There are some additional characters for punctuation, etc. in the Bengali character set, which are ignorable for collation. The complete encoded character set in Unicode is given in [4].

3.1.2. Script Details

3.1.2.1. Consonants and Vowels

Bengali is written from left to right. Space is used to mark word boundaries. Letters are uncased and are grouped together based on place and manner of articulation. The characters in Bengali script hang from a horizontal line, called the head stroke. When writing, these characters within a word head strokes merge to form single base line, as shown for the word BAABAA (father) in Figure 3.4.

ব (Letter Ba) + ◌া (Vowel AA) + ব (Letter Ba) + ◌া (Vowel AA) = বাবা

Figure 3.4. Merging of Head Strokes of Bengali Characters

The consonants in Bengali have an inherent [ɔ][1] sound by default. For example Bengali letter ক represents [kɔ] and not [k] sound. Virama is placed below delete the vowel sound and get the pure consonantal sound. However, the use of Virama is often implied and optionally written by Bangla speakers.

The vowels take the independent vowel shape if they are in a syllable without an onset consonant. In case they are in a syllable with an onset consonant, they attach with the consonant taking the dependent shape. Thus, all vowels have an independent and dependent shape, except the vowel [ɔ] which only has an independent shape অ. It does not have a dependent shape as it is inherently present with each consonant by default if not explicity deleted by Virama or over-ridden by another dependent vowel. The dependent vowels attach at the front, back, top or bottom of a consonant. These are illustrated in Table 3.2. In some cases the vowel splits into two halves and is placed across consonant such that one half is at right while other is at left.

Table 3.2. Dependent Vowels with Consonant [k]

Consonant + Dependent Vowel	Joined Form	Comment
ক+◌ি	কি	Connects to left of consonant
ক+◌ী	কী	Connects to right of consonant
ক+◌ু	কু	Connects to base of consonant
ক+◌ৌ	কৌ	Wraps around the consonant

As is shown in Table 3.2, the vowel is typed after the consonant no matter where it attaches. Also, only one vowel can connect to a consonant at a time. The dependent vowels can not occur with independent vowels or by themselves.

3.1.2.2. Conjunct Consonants

In Bengali two or more consonants may join together to form complex conjuncts with alternate shapes. In Unicode, Virama is placed on the first consonant in a pair to enforce the conjoined shape of the consonants [4]. Some conjuncts and non-conjunct shapes are given in Table 3.3.

[1] Square brackets [ ] are conventionally used to represent a phone or a sound.

Like other Indic languages, র (or /r/) also forms different shapes in consonant clusters. When in initial position it is displayed as a mark to top, and when at the end it appears as a wavy line below the consonant to which it connects [4], as shown in last two rows of Table 3.3.

Table 3.3. Conjunct Consonants

Consonants **C1 C2**	**Clustered Form** **C1 + C2**	**Conjunct Form** **C1 + ◌্ + C2**
ক ষ	কষ	ক্ষ
স ক	সক	স্ক
ব দ	বদ	ব্দ
র য	রয	র‍্য
ব র	বর	ব্র

A more comprehensive list of conjunct consonants can be viewed at [14].

3.2. Collation

Bengali collation sequence has been defined by Bangla Academy, the language authority of Bangladesh. This section elaborates on this collation sequence for Bengali and an algorithmic implementation using UCA [2] for Bengali collation.

In Bengali all characters have primary level significance for collation purposes. Numerals and currency symbols are given smallest weight; these are followed by independent vowels, modifiers, consonants and dependent vowels. However, before collation is applied some text processing is required. Details of the text processing are also presented.

3.2.1. Text Processing

3.2.1.1. Reordering

As mentioned above the independent vowels combine with consonants in different manners i.e. joining to right, left, above or below. In hand-written orthography, old type-writers and non-standard Bengali encodings, the vowels that attach to the left are written first followed by a consonant. Others are written after the consonant. Thus, the typing order for কৈ is ◌ৈ + ক and for

কী is ক + ◌ী. For collation, the logical comparison order is consonant and then the dependent vowel, wherever the vowels attaches to the consonant. The typing sequence just discussed is inconsistent and thus the logical comparison between two combinations is not possible. Thus, the preceding vowel needs to be re-ordered, after the consonant, if a comparison has to be enabled. This is true for all the encodings which require dependent vowels to be typed before the consonant. However, the Unicode standard for Bengali requires consonant + vowel typing order whether the vowel visually appears after or before the consonant. The visual placement is separately handled in the rendering process. Therefore, if the Unicode encoding is followed, no reordering is required.

3.2.1.2. Normalization

There are different ways some Bengali characters, both consonants and vowels, can be encoded in Unicode. Thus, normalization is required before collation can be done. As discussed during the general discussion on collation in the second chapter, both composed or decomposed forms may be taken to do the collation, as long as it is consistently done. This section lists some of the equivalent forms for Bengali.

The first set of equivalents in Bengali are formed due encoding of Nukta as a combining character ◌়. (U+09BC). Nukta combines with consonants to give additional consonants, which are also separately encoded. Examples are given in Table 3.4. below.

Table 3.4. Normalization Due to Nukta in Bengali [4]

Decomopsed Form	Unicodes of Decomposed Form	Equivalent Composed Form	Unicode of Composed Form
ড ◌়	09A1 09BC	ড়	09DC
ঢ ◌়	09A2 09BC	ঢ়	09DD
য ◌়	09AF 09BC	য়	09DF

Similarly, dependent vowels which have two parts and surround the consonant also have equivalent encodings, equivalent to the case where a single vowel is split into the parts which come before and after the consonant respectively. The equivalents are given in Table 3.5. As can be seen in the table, both forms render in the same way when combined with a consonant are equivalent in terms of collation.

Table 3.5. Normalization Due to Glyph Splitting of Two-Part Dependent Vowels

Decomposed Form	Unicodes of Decomposed Form	Use with a Consonant	Equivalent Composed Form	Unicode of Composed Form	Use with a Consonant
◌ে ◌া	09CB 09BE	ক ◌ে ◌া = কো	◌ো	09CB	ক ◌ো = কো
◌ে ◌ৗ	09C7 09D7	ক ◌ে ◌ৗ = কৌ	◌ৌ	09CC	ক ◌ৌ = কৌ

One can form half shape of consonants in Indic scripts. Unicode enables that by typing Virama after the consonant. In a special case, Bengali conjunct character 'tta' can be encoded in multiple ways, but must show the same behavior for collation. Thus, the variations must be normalized to represent the same collation weight.

Table 3.6. Encoding and Rendering Variations of 'tta' Conjunct with Khanda Ta Character

Constituent Characters	Unicode Sequence	Rendered Variant Form
ত ◌্ ত	09A4 09CD 09A4	ত্ত
ত ◌্ ZWJ ত	09A4 09CD 200D 09A4	ৎত
ত ◌্ ZWNJ ত	09A4 09CD 200C 09A4	ত্‌ত
ৎ ত	09CE 09A4	ৎত

The normalization with Khanda Ta is different from the first two cases discussed because the final conjunct form is not encoded. Thus, the sequence can only be equated in decomposed forms and cannot be mapped onto a single composed form.

3.2.1.3. 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, presented in Tables 3.4 and 3.5, is illustrated in Table 3.7.

Table 3.7. Contraction to Single Collation Element from Multiple Encoded Characters

Glyph	Unicodes of Decomposed Form	Unicode of Composed Form	Collation Element	Unicode Name
ড ◌়	09A1 09BC	09DC	15BD 0020 0002	LETTER RRA
ঢ ◌়	09A2 09BC	09DD	15BF 0020 0002	LETTER RHA
য ◌়	09AF 09BC	09DF	15CC 0020 0002	LETTER YYA
◌ে ◌ৗ	09C7 09D7	09CC	15E3 0020 0002	VOWEL SIGN AU
◌ে ◌া	09C7 09BE	09CB	15E2 0020 0002	VOWEL SIGN O

3.2.1.4. Conjunct Consonants

The formation of alternate glyphs for conjuncts does not change input sequence logically but only visually. Collation is dependent on the logical sequence and thus is not affected by the change in shape. The Zero Width Joiner and Zero Width Non-Joiner are ignored in the process. However, ambiguity occurs in case of the combination of Ra and Ya, where Zero Width Non-Joiner plays a significant role. See [26] for further details.

3.2.2. Collation Elements

In order to realize Bengali collation as defined by Bangla Academy [15], following collation element table may be used. The table gives multiple entries in relevant columns if required. The table is further divided into sub-sections for various families of characters, including signs, numerals, dependent vowels, characters and dependent vowels.

Table 3.8. Collation Elements for Bengali Language

Glyph	Unicode	Collation Elements	Unicode Name
← Various Signs →			

◌়	09BC	13A0 0020 0002	BENGALI SIGN NUKTA
◌ং	0982	13A2 0020 0002	BENGALI SIGN ANUSVARA
◌ঃ	0983	13A3 0020 0002	BENGALI SIGN VISARGA
◌ঁ	0981	13A4 0020 0002	BENGALI SIGN CANDRABINDU
← Numerals & Currency Symbols →			
৸	09F8	0DC7 0020 0002	BENGALI CURRENCY NUMERATOR ONE LESS THAN THE DENOMINATOR
৹	09F9	0DC8 0020 0002	BENGALI CURRENCY DENOMINATOR SIXTEEN
৺	09FA	0350 0020 0002	BENGALI ISHAR
৲	09F2	0E12 0020 0002	BENGALI RUPEE MARK
৳	09F3	0E13 0020 0002	BENGALI RUPEE SIGN
০	09E6	0E29 0020 0002	BENGALI DIGIT ZERO
১	09E7	0E2A 0020 0002	BENGALI DIGIT ONE
৴	09F4	0E2A 0020 0002	BENGALI CURRENCY NUMERATOR ONE
২	09E8	0E2B 0020 0002	BENGALI DIGIT TWO
৵	09F5	0E2B 0020 0002	BENGALI CURRENCY NUMERATOR TWO
৩	09E9	0E2C 0020 0002	BENGALI DIGIT THREE
৶	09F6	0E2C 0020 0002	BENGALI CURRENCY NUMERATOR THREE
৪	09EA	0E2D 0020 0002	BENGALI DIGIT FOUR
৷	09F7	0E2D 0020 0002	BENGALI CURRENCY NUMERATOR FOUR
৫	09EB	0E2E 0020 0002	BENGALI DIGIT FIVE

৬	09EC	0E2F 0020 0002	BENGALI DIGIT SIX
৭	09ED	0E30 0020 0002	BENGALI DIGIT SEVEN
৮	09EE	0E31 0020 0002	BENGALI DIGIT EIGHT
৯	09EF	0E32 0020 0002	BENGALI DIGIT NINE
← Independent Vowels →			
অ	0985	12A2 0020 0002	BENGALI LETTER A
আ	0986	12A3 0020 0002	BENGALI LETTER AA
ই	0987	12A4 0020 0002	BENGALI LETTER I
ঈ	0988	12A5 0020 0002	BENGALI LETTER II
উ	0989	12A6 0020 0002	BENGALI LETTER U
ঊ	098A	12A7 0020 0002	BENGALI LETTER UU
ঋ	098B	12A8 0020 0002	BENGALI LETTER VOCALIC R
ৠ	09E0	12A9 0020 0002	BENGALI LETTER VOCALIC RR
ঌ	098C	12AA 0020 0002	BENGALI LETTER VOCALIC L
ৡ	09E1	12AB 0020 0002	BENGALI LETTER VOCALIC LL
এ	098F	12AC 0020 0002	BENGALI LETTER E
ঐ	0990	12AD 0020 0002	BENGALI LETTER AI
ও	0993	12AE 0020 0002	BENGALI LETTER O
ঔ	0994	12AF 0020 0002	BENGALI LETTER AU
← Consonants →			
ক	0995	15B0 0020 0002	BENGALI LETTER KA

খ	0996	15B1 0020 0002	BENGALI LETTER KHA
গ	0997	15B2 0020 0002	BENGALI LETTER GA
ঘ	0998	15B3 0020 0002	BENGALI LETTER GHA
ঙ	0999	15B4 0020 0002	BENGALI LETTER NGA
চ	099A	15B5 0020 0002	BENGALI LETTER CA
ছ	099B	15B6 0020 0002	BENGALI LETTER CHA
জ	099C	15B7 0020 0002	BENGALI LETTER JA
ঝ	099D	15B8 0020 0002	BENGALI LETTER JHA
ঞ	099E	15B9 0020 0002	BENGALI LETTER NYA
ট	099F	15BA 0020 0002	BENGALI LETTER TTA
ঠ	09A0	15BB 0020 0002	BENGALI LETTER TTHA
ড	09A1	15BC 0020 0002	BENGALI LETTER DDA
ড়	09DC	15BD 0020 0002	BENGALI LETTER RRA
ড ◌়	09A1 09BC	15BD 0020 0002	BENGALI LETTER RRA
ঢ	09A2	15BE 0020 0002	BENGALI LETTER DDHA
ঢ়	09DD	15BF 0020 0002	BENGALI LETTER RHA
ঢ ◌়	09A2 09BC	15BF 0020 0002	BENGALI LETTER RHA
ণ	09A3	15C0 0020 0002	BENGALI LETTER NNA
ত	09A4	15C1 0020 0002	BENGALI LETTER TA
থ	09A5	15C2 0020 0002	BENGALI LETTER THA
দ	09A6	15C3 0020 0002	BENGALI LETTER DA

ধ	09A7	15C4 0020 0002	BENGALI LETTER DHA
ন	09A8	15C5 0020 0002	BENGALI LETTER NA
প	09AA	15C6 0020 0002	BENGALI LETTER PA
ফ	09AB	15C7 0020 0002	BENGALI LETTER PHA
ব	09AC	15C8 0020 0002	BENGALI LETTER BA
ভ	09AD	15C9 0020 0002	BENGALI LETTER BHA
ম	09AE	15CA 0020 0002	BENGALI LETTER MA
য	09AF	15CB 0020 0002	BENGALI LETTER YA
য় য ◌়	09DF 09AF 09BC	15CC 0020 0002 15CC 0020 0002	BENGALI LETTER YYA BENGALI LETTER YYA
র	09B0	15CD 0020 0002	BENGALI LETTER RA
ৰ	09F0	15CE 0020 0002	BENGALI LETTER RA WITH MIDLE DIAGONAL
ল	09B2	15CF 0020 0002	BENGALI LETTER LA
ৱ	09F1	15D0 0020 0002	BENGALI LETTER RA WITH LOWER DIAGONAL
শ	09B6	15D1 0020 0002	BENGALI LETTER SHA
ষ	09B7	15D2 0020 0002	BENGALI LETTER SSA
স	09B8	15D3 0020 0002	BENGALI LETTER SA
হ	09B9	15D4 0020 0002	BENGALI LETTER HA
ঽ	09BD	15D5 0020 0002	BENGALI SIGN AVAGRAHA
ৎ	09CE	[15C1 0020 0002],[15E4 0020 0002]	BENGALI LETTER KHANDA TA

← Dependant Vowels →			
◌া	09BE	15D6 0020 0002	BENGALI VOWEL SIGN AA
◌ি	09BF	15D7 0020 0002	BENGAL VOWEL SIGN I
◌ী	09C0	15D8 0020 0002	BENGAL VOWEL SIGN II
◌ু	09C1	15D9 0020 0002	BENGAL VOWEL SIGN U
◌ূ	09C2	15DA 0020 0002	BENGAL VOWEL SIGN UU
◌ৃ	09C3	15DB 0020 0002	BENGAL VOWEL SIGN VOCALIC R
◌ৄ	09C4	15DC 0020 0002	BENGAL VOWEL SIGN VOCALIC RR
◌ৢ	09E2	15DD 0020 0002	BENGAL VOWEL SIGN VOCALIC L
◌ৣ	09E3	15DF 0020 0002	BENGAL VOWEL SIGN VOCALIC LL
◌ে	09C7	15E0 0020 0002	BENGAL VOWEL SIGN E
◌ৈ	09C8	15E1 0020 0002	BENGAL VOWEL SIGN AI
◌ো	09CB	15E2 0020 0002	BENGAL VOWEL SIGN O
◌ে ◌া	09C7 09BE	15E2 0020 0002	BENGAL VOWEL SIGN O
◌ৌ	09CC	15E3 0020 0002	BENGAL VOWEL SIGN AU
◌ে ◌ৗ	09C7 09D7	15E3 0020 0002	BENGAL VOWEL SIGN AU
◌্	09CD	15E4 0020 0002	BENGALI SIGN VIRMA
◌ৗ	09D7	15E5 0020 0002	BENGALI AU LENGTH MARK

3.2.3. Results

Table 3.9 shows output obtained by sorting a sample input using the collation elements given in Table 3.8.

Table 3.9. Input and Corresponding Sorted Output for Bengali

Input		Output	
ঋতু	এ১	অংশ	এও
কোল৪	ঔৎকষ	অংশাংশ	এঃ
ইঃ	ক১	অংশাংশি	এঁঠড়
অকথিত	কই১	অংশানো	ঐ১
কৌচ	এঁঠড়	অংশী	ও২
অকুণ্ঠ	অংশ	অংশে	ওঁ
ইউনিফম	কওম	অকথিত	ঔৎকষ
ইংকার	কতক	অকুণ্ঠ	ঔৎসুক্
অংশী	ওঁ	ইউনানি	ক১
ইঁচড়	কেল	ইউনিফম	ক৪
উওল	অংশাংশ	ইংকার	কই১
উদার	কোল১	ইঃ	কওম
ঊঢ়	অংশে	ইঁচড়	কওলা
এও	ক৪	উওল	কত
ঋক্	ঔৎসুক্	উদার	কতক
অংশাংশি	কতকটা	ঊঢ়	কতকটা
অংশাংনো	কত	ঋক্	কেল
এ২	ঐ১	ঋতু	কোল১

এইতো	কওলা	এ১	কোল৪
ইউনানি	ও২	এ২	কৌচ
এঃ	কৌচ	এইতো	কৌচ

3.3. Conclusion

Bengali, like other Indic languages, has single level of collation. All characters are sorted at primary level with numerals and currency symbols, independent vowels, modifiers, consonants and dependent vowels sorted in this order. The sorting 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.

4. Dzongkha

Dzongkha is a Sino-Tibetan language related to Tibetan. It has 0.13 million first-language speakers [28] and approximately 0.5 Million total speakers [30] in Bhutan. Dzongkha is the native language of eight western districts of Bhutan (Thimphu, Paro, Punakha, Wangdue, Phodrang, Gasa, Ha, Dhakana, and Chukha) and also recognized as the national and official language of the country. Dzongkha speakers also reside in India (specifically West Bengal) and Nepal [29].

4.1. Writing System

4.1.1. Character Set

Dzongkha is written in Tibetan script which itself is motivated by the syllabic Devanagari writing system. Dzongkha character set consists of 30 consonants and four vowels [30]. Each consonant has inherent /a/ sound. The consonants are given below, arranged according to place and manner of articulation.

ཀ ཁ ག ང ཅ ཆ ཇ ཉ ཏ ཐ ད ན པ ཕ བ མ ཙ ཚ ཛ ཝ

ཞ ཟ འ ཡ ར ལ ཤ ས ཧ ཨ

Figure 4.1. Dzongkha Consonants

Six extra consonants are also used in Dzongkha, which were originally to write Sanskrit loan words, but now they are also used to write other foreign words. These are also known as reversed letters, because some are mirror images of the consonants above [31]. Figure below shows these characters.

ཊ ཋ ཌ ཎ ཥ ཀྵ

Figure 4.2. Reversed Consonants of Dzongkha

Dzongkha writing system has a single independent vowel ཨ or /a/. In addition, there are four dependent vowels shown below (/i, u, e, o/ respectively), which combine with consonants.

◌ི ◌ུ ◌ེ ◌ོ

Figure 4.3. Vowels of Dzongkha

Additional special characters and modifier symbols are used in Dzongkha. These are shown below.

Table 4.1. Dzongkha Special Characters

Name	Glyph	Usage
Sign Rnam Bcad	◌ཿ	ཨཿ
Sign Sna Ldan	◌ྃ	ཨྃ
Mark Tsheg	་	ཀེ་ནེ་ར་
Mark Shad	།	
Mark Nyis Shad	༎	

Sign Rnam Bcad, also known as Visarga, adds voiceless breath after the consonant and is generally used to write Sanskrit words. Sign Sna Ldan is same as Anusvara in Indic languages, and is used to nasalize the vowel and is also used for Sanskrit words [32]. Tsheg is used to separate character clusters or units, roughly but not exactly the same as a syllable (former is glyph/grapheme motivated, whereas latter is phonological motivated).

Mark Shad represents end of an expression while Nyis Shad marks a change in topic. These are commonly used punctuation mark. These are derived from Devanagri Danda and Double Danda and are “roughly equivalent to the comma and period” [5].

Dzongkha has its own numerals. These are shown below.

༠ ༡ ༢ ༣ ༤ ༥ ༦ ༧ ༨ ༩

Figure 4.4. Dzongkha Numerals

Half form for each Dzongkha digit also exists, as is shown below [35].

༳ ༪ ༫ ༬ ༭ ༮ ༯ ༰ ༱ ༲

Figure 4.5. Numerals with Half Forms in Dzongkha

4.1.2. Script Details

Dzongkha is written from left to right. The written form has multiple level of stacking of consonants and vowels. Dzongkha does not use regular spaces between words. Text flows in a continuum. Space is used rarely, and may not mark a boundary in the writing system. As mentioned already punctuation marks Danda and Double Danda are used to mark expression and topic boundaries. Unlike other South East Asian languages syllable-like character clusters are explicitly separated by special mark Tsheg. Detailed syllable structure is discussed in the next section.

4.1.2.1. Syllable Structure

A syllable in Dzongkha can have one to six characters including one to four consonant characters [31]. Each syllable has exactly one core consonant character known as 'root' or 'radical'. A syllable can optionally have a prefix, a suffix, a subscribed or a super-scribed letter and vowels (dependent and independent). The figure below shows template syllable structure. This template is explained below in context of the information available at [31].

	$V_{0,1}$		
	$Sp_{0,1}$		$V_{0,1}$
$P_{0,1}$	$\mathbf{R_1}$	$Sf_{0,2}$	ཨ$_{0,1}$
	$Sb_{0,1}$		$V_{0,1}$
	$V_{0,1}$		

Figure 4.6. Generic Syllable Structure

- 'R' represents the root, and subscript digit '1' represents that this root may be exactly one letter in a syllable.
- 'P' represents prefix. Prefix is unpronounced in most cases but it does modify the pronunciation of the following root in a few cases. Each syllable can have at most one prefix. The five letters which can occur in this position are given below.

ག ད བ མ འ

Figure 4.7. Dzongkha Prefix Letters

- 'Sf' represents suffix characters. Suffix marks the end of a syllable if independent vowel ཨ does not follow. A syllable can have up to two suffixes. Second suffix is known as secondary suffix. There are ten consonants that can take role of primary suffix (shown below). Only two consonants ས (Letter SA; U+0F66) and ད (Letter DA; U+0F51) can act as a secondary suffix. In some cases, primary suffix cancels the inherit /a/ sound of radical by adding its own sound. In other cases, it modifies the vowel of root letter. Sometimes it can have both these impacts.

ག ང ད ན བ མ འ ར ལ ས

Figure 4.8. Dzongkha Suffix Letters

- ‘Sp’ represents super-scribed letters. These are placed at the top of other consonants. Super-scribed letters modify the pronunciation of their host consonants by raising its tone or pitch. The three super-scribed letters when attach with different consonants to form three categories of letters namely Ra-go, La-go and Sa-go letters. These are shown below along with all the consonants they can attach with.

Table 4.2. Super-Scribed Letters

	Letter	Usage
Ra-Go	ར	རྐ རྒ རྔ རྗ རྙ རྟ རྡ རྣ རྦ རྨ རྩ རྫ
La-Go	ལ	ལྐ ལྒ ལྔ ལྕ ལྗ ལྟ ལྡ ལྤ ལྦ ལྷ
Sa-Go	ས	སྐ སྒ སྔ སྙ སྟ སྡ སྣ སྤ སྦ སྨ སྩ

- ‘Sb’ represents sub-scribed letters. These are placed underneath other consonants. The four sub-scribed letters when attach with different consonants form four categories of letters namely Ya-ta, Ra-ta, La-ta and Wa-zur letters. These are shown below along with the consonants they attach with. Like super-scribed letters these also modify the sound of the consonant that they attach with. Wa-zur combination does not modify the sound of the host consonant.

Table 4.3. Dzongkha Sub-Scribed Letters

	Letter	Usage
Ya-Ta	ཡ	ཀྱ ཁྱ གྱ པྱ ཕྱ བྱ མྱ ཧྱ
Ra-Ta	ར	ཀྲ ཁྲ གྲ ཏྲ ཐྲ དྲ ནྲ པྲ ཕྲ བྲ སྲ མྲ ཧྲ
La-Ta	ལ	ཀླ གླ བླ རླ སླ ཟླ
Wa-Zur	ཝ	ཀྭ ཁྭ གྭ ཉྭ དྭ ཙྭ ཚྭ ཞྭ ཟྭ རྭ ལྭ ཤྭ ཧྭ

The shapes of all sub-scribed letters change except for La-ta, when they attach underneath another consonant.

- 'V' represents dependent vowels (shown above in Figure 4.3). When they attach, the dependent vowels replace the inherent vowel associated with consonants.

Table 4.4. Dependent Vowels with Consonant Ka

C + V	Joined Form	Sound	Comment
ཀ	ཀ	[ka]	Inherent Vowel
ཀ+◌ོ	ཀོ	[ko]	Vowel connects at its Top
ཀ+◌ུ	ཀུ	[ku]	Vowel connects to its Base

Only the root consonant R or independent vowel འ in a syllable can take dependent vowels.

- The consonants in Dzongkha have inherent sound /a/ by default. For example letter ཀ possesses sound [ka]. Unlike Indic languages Virama is not used in Dzongkha to nullify the /a/ sound. Only the root consonant in a syllable has inherent /a/ sound which can be changed through four dependent vowels. For example syllable དག has sound [dag] and not [Daga]. In this case suffix ག marks the end of syllable. If the syllable has a consonant cluster in the onset and has to end with inherent sound /a/ then it should end with independent vowel འ. So the syllable དགའ has sound [dga] [5]. The vowel འ can also take independent vowels to change the syllable sound to [dgi, dgu] etc.

More details on each of these can be found at [31].

The pronunciation of a character cluster is based on its root letter. So if a prefix or a suffix letter is mistakenly identified as root the reader would pronounce the word incorrectly. This frequently happens to new learners. Root can be identified through following five rules [31].

- Only root in a syllable can take vowels, except for in case of phase connector འི and independent vowel འ

- Only root letter can have sub or super-scribed letters
- A two letter syllable with no vowels has first letter as root. Second letter is suffix
- A three letter syllable usually has middle letter as a root. However in presence of secondary suffix any of the first or second character can be the root
- A four letter syllable always has second letter as root

The following word བརྐོས་སྔེར has two character clusters བརྐོས and སྔེར as shown in the figure below.

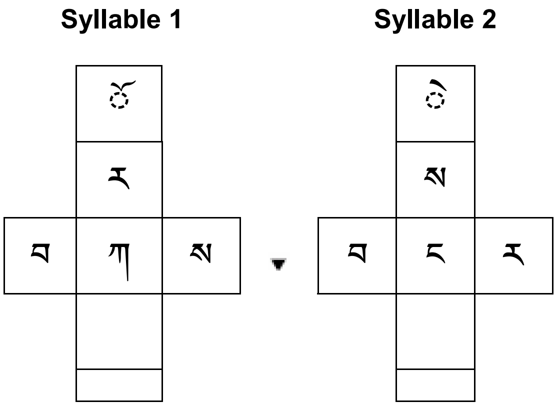

Figure 4.9. Sample Syllables

4.1.2.2. *Conjunct Consonants*

Dzongkha frequently forms a large number of conjunct consonants with sub and super-scribed consonants. Three of these are shown below.

Table 4.5. Conjunct Consonants [29]

Characters	Conjunct Form
ར ◌ྐ ◌ྭ	རྐྭ
ས ◌ྐ ◌ྱ	སྐྱ
ར ◌ླ	རླ

All the consonants in Dzongkha are encoded in two forms. The first form is for the original form of consonants when they appear as normal form or as a top element of a conjunct. Second form is the sub-joined form of each consonant. The sub-joined forms can appear in conjunct anywhere but at the top [5] and are used because Virama or Halanta is not explicitly used. For example ས and ར in the above figure are normal forms while ◌ྐ, ◌ྑ, ◌ྒ, ◌ྲ and ◌ྱ are sub-joined forms.

4.2. *Collation*

Dzongkha is mainly sorted at primary level. However a few special marks and that are given secondary level weights and the reversed characters used (see figure 4.2) to represent foreign words are given both primary and tertiary weights. For example ཏ and ཊ exhibit case level difference. These are same at primary level but differ at tertiary level [36]. At primary level numerals and their half forms are given smallest weight; these are followed by consonants and their cluster variants, followed by special modifiers. These are followed by vowels and finally sub-joined forms of consonants. Further details can be viewed in collation element table given below in Table 4.8.

4.2.1. Text Processing

4.2.1.1. Syllabification

Dzongkha like Lao is collated syllable by syllable. Each syllable in a string is compared with its corresponding syllable in the other string. Second and subsequent syllables are only compared when there is previous syllables are identical. However the case of Dzongkha is less complex as the syllable boundaries are explicitly defined by Tseg mark and the collation process is not required to detect these automatically. This inter-syllabic mark Tseg has been assigned lightest weight which ensures syllable by syllable comparison; alternatively, it could also be ignored if more explicit syllable separation process is introduced.

4.2.1.2. Normalization

In Dzongkha consonants conjoin with other consonants and vowels to form complex conjuncts. Unicode consortium has assigned code points to few such frequently used conjuncts. As a result such ligatures/conjuncts can be obtained in two ways; either by typing sequence of its constituent characters or by inserting the code point of the conjunct. For example ◌ཷ (U+0F77) can also be obtained by typing sequence ◌ྲ (U+0FB2) + ◌ཱྀ (U+0F81). The cluster ◌ཱྀ can further be decomposed into ◌ཱ (U+0F71) + ◌ྀ (U + 0F80). So the cluster ◌ཷ has two other equivalent forms. For collation these three are essentially the same. Therefore these are required to be normalized to one of the three forms. The cluster ◌ཱི however is not a linguistic entity. So it is required to be broken into its constituents in order to obtain proper results. Other such clusters are shown in the figure below.

Table 4.6. Normalization Cases

Decomposed Form	Unicode of Decomposed Form	Equivalent Composed Form	Unicode of Composed Form
ཀ ◌ྵ	0F40 0FB5	ཀྵ	0F69
◌ཱ ◌ི	0F71 0F72	◌ཱི	0F73
◌ྡ ◌ྷ	0FA1 0FB7	◌ྡྷ	0FA2
◌ྦ ◌ྷ	0FA6 0FB7	◌ྦྷ	0FA7
◌ླ ◌ྀ	0FB3 0F80	◌ླྀ	0F78
◌ྲ ◌ཱ ◌ྀ	0FB2 0F71 0F80	◌ཷ	0F77
◌ཱ ◌ུ	0F71 0F74	◌ཱུ	0F75
ཛ ◌ྷ	0F5B 0FB7	ཛྷ	0F5C
བ ◌ྷ	0F56 0FB7	བྷ	0F57
ད ◌ྷ	0F51 0FB7	དྷ	0F52
◌ྲ ◌ྰ	0FB2 0FB0	◌ྲྀ	0F76
◌ླ ◌ཱ ◌ྀ	0FB3 0F71 0F80	◌ཹ	0F79
◌ཱ ◌ྀ	0F71 0F80	◌ཱྀ	0F81
◌ྒ ◌ྷ	0F92 0FB7	◌ྒྷ	0F93
◌ྫ ◌ྷ	0FAB 0FB7	◌ྫྷ	0FAC
◌ྜ ◌ྷ	0F9C 0FB7	◌ྜྷ	0F9D
◌ྐ ◌ྵ	0F90 0FB5	◌ྐྵ	0FB9
ཌ ◌ྷ	0F4C 0FB7	ཌྷ	0F4D

ག ◌ྷ	0F42 0FB7	གྷ	0F43

4.2.1.3. Contraction

Contraction is a very important phenomenon for Dzongkha. A root can optionally have prefix, suffix, sub- and super-joined consonants which combine with root to form different collation entities. A Prefix + Root combination, for instance, is slightly heavier than the root itself. Consider root letter ཀ, which has the following cluster variants: དཀ, བཀ, རྐ, ལྐ, སྐ, བརྐ, བསྐ, each having its own collation identity. Not all the main consonants have these variant forms so a generic rule can not be defined. For 30 consonants there are 133 such clusters. These clusters are required to be mapped onto single collation element. Below is list of few such formations. Detailed list along with collation elements can be viewed in Table 4.8.

Table 4.7. Contraction Cases

Glyph	Unicodes for Contraction
ད + ཀ = དཀ	0F51 + 0F40
བ + ཀ = བཀ	0F56 + 0F40
ར + ◌ྐ = རྐ	0F62 + 0F90
ལ + ◌ྐ = ལྐ	0F63 + 0F90
ས + ◌ྐ = སྐ	0F66 + 0F90
བ + ར + ◌ྐ = བརྐ	0F56 + 0F62 + 0F90
བ + ས + ◌ྐ = བསྐ	0F56 + 0F66 + 0F90

4.2.1.4. Context Sensitive Collation Element Assignment

As mentioned before, collation in Dzongkha is based on syllables. Within each syllable, the main consonant i.e. root dictates where and how it sorts. It is therefore important to classify whether a character in a syllable is playing root or prefix. There are special cases in Dzongkha where a third character or even fourth character is required to find out the root and prefix/suffix in a syllable. Consider following example of དག. Without a third character it is hard to determine root out of this sequence. In case of དགར, letter ག is the root letter, so དགར sorts under letter ག. However, in case of གདག, letter ད is the root letter so གདག sorts under ད. In the former sequence ད is the prefix, while in latter it is the root.

Such ambiguities are by no means rare in Dzongkha and these are very hard for collation process to detect. More complex case is of དཕ where a fourth code point is also required to determine the root letter. Given below is a comprehensive list of such ambiguities. The collation elements proposed in this chapter does not tackle this problem.

དག བག མག འག དང མང དབ འབ དམ གད བད མད འད
གན མན བར གས བས

Figure 4.10. Ambiguous Cases [33]

4.2.1.5. Reordering

In Dzongkha text flows from left to right as well as in vertical direction. Multiple consonants and vowels conjoin together to form character stacking. Consider following example: སྒྲི. There has to be a unique standardized input sequence to obtain this conjunct. The order of characters in input sequence is important for collation. Vowel should logically be after consonants (normal or sub-joined) to obtain correct sorting results.

The Unicode standard for Dzongkha recommends normal form of consonant + subjoined forms + vowel, typing order. This is the required order. So reordering is not explicitly tackled during collation process.

4.2.1.6. Conjunct Consonants

The formation of conjuncts is a visual process and does not change the input sequence logically. Therefore conjoining characters have no bearing on collation.

4.2.2. Unicode Collation Elements

The collation elements for Dzongkha characters are defined below. This order is observed by Dzongkha dictionary approved from Dzongkha language authority [34].

Table 4.8. Collation Elements

Glyph	Unicode	Collation Elements	Unicode Name
ཀ	0F40	1375 0020 0002	TIBETAN LETTER KA
དཀ	0F51 0F40	1376 0020 0002	TIBETAN LETTER DA + TIBETAN LETTER KA
བཀ	0F56 0F40	1378 0020 0002	TIBETAN LETTER BA + TIBETAN LETTER KA
རྐ	0F62 0F90	1379 0020 0002	TIBETAN LETTER RA + TIBETAN SUBJOINED LETTER KA
ལྐ	0F63 0F90	137B 0020 0002	TIBETAN LETTER LA +TIBETAN SUBJOINED LETTER KA
སྐ	0F66 0F90	137C 0020 0002	TIBETAN LETTER SA + TIBETAN SUBJOINED LETTER KA
བརྐ	0F56 0F62 0F90	137F 0020 0002	TIBETAN LETTER BA + TIBETAN LETTER RA + TIBETAN SUBJOINED LETTER KA
བསྐ	0F56 0F66 0F90	1380 0020 0002	TIBETAN LETTER BA + TIBETAN LETTER SA + TIBETAN SUBJOINED LETTER KA
ཁ	0F41	1382 0020 0002	TIBETAN LETTER KHA
མཁ	0F58 0F41	1383 0020 0002	TIBETAN LETTER MA + TIBETAN LETTER KHA
འཁ	0F60 0F41	1385 0020 0002	TIBETAN LETTER –A + TIBETAN LETTER KHA
ག	0F42	1386 0020 0002	TIBETAN LETTER GA
དག	0F51 0F42	1388 0020 0002	TIBETAN LETTER DA + TIBETAN LETTER GA
བག	0F56 0F42	1389 0020 0002	TIBETAN LETTER BA + TIBETAN LETTER GA
མག	0F58 0F42	138B 0020 0002	TIBETAN LETTER MA + TIBETAN LETTER GA
འག	0F60 0F42	138C 0020 0002	TIBETAN LETTER –A + TIBETAN LETTER GA
རྒ	0F62 0F92	138F 0020 0002	TIBETAN LETTER RA + TIBETAN SUBJOINED LETTER GA
ལྒ	0F63 0F92	1390 0020 0002	TIBETAN LETTER LA + TIBETAN SUBJOINED LETTER GA
སྒ	0F66 0F92	1392 0020 0002	TIBETAN LETTER SA + TIBETAN SUBJOINED LETTER GA

བརྒ	0F56 0F62 0F92	1393 0020 0002	TIBETAN LETTER BA + TIBETAN LETTER RA + TIBETAN SUBJOINED LETTER GA
བསྒ	0F56 0F66 0F92	1395 0020 0002	TIBETAN LETTER BA +TIBETAN LETTER SA + TIBETAN SUBJOINED LETTER GA
ང	0F44	1396 0020 0002	TIBETAN LETTER NGA
དང	0F51 0F44	1398 0020 0002	TIBETAN LETTER DA +TIBETAN LETTER NGA
མང	0F58 0F44	1399 0020 0002	TIBETAN LETTER MA + TIBETAN LETTER NGA
རྔ	0F62 0F94	139A 0020 0002	TIBETAN LETTER RA + TIBETAN SUBJOINED LETTER NGA
ལྔ	0F63 0F94	139B 0020 0002	TIBETAN LETTER LA + TIBETAN SUBJOINED LETTER NGA
སྔ	0F66 0F94	139C 0020 0002	TIBETAN LETTER SA + TIBETAN SUBJOINED LETTER NGA
བརྔ	0F56 0F62 0F94	139D 0020 0002	TIBETAN LETTER BA + TIBETAN LETTER RA + TIBETAN SUBJOINED LETTER NGA
བསྔ	0F56 0F66 0F94	139F 0020 0002	TIBETAN LETTER BA+ TIBETAN LETTER SA +TIBETAN SUBJOINED LETTER NGA
ཅ	0F45	13A0 0020 0002	TIBETAN LETTER CA
གཅ	0F42 0F45	13A2 0020 0002	TIBETAN LETTER GA + TIBETAN LETTER CA
བཅ	0F56 0F45	13A3 0020 0002	TIBETAN LETTER BA + TIBETAN LETTER CA
ལྕ	0F63 0F95	13A5 0020 0002	TIBETAN LETTER LA + TIBETAN SUBJOINED LETTER CA
བལྕ	0F56 0F63 0F95	13A6 0020 0002	TIBETAN LETTER BA+ TIBETAN LETTER LA+ TIBETAN SUBJOINED LETTER CA
ཆ	0F46	13A8 0020 0002	TIBETAN LETTER CHA
མཆ	0F58 0F46	13A9 0020 0002	TIBETAN LETTER MA + TIBETAN LETTER CHA
འཆ	0F60 0F46	13AB 0020 0002	TIBETAN LETTER –A + TIBETAN LETTER CHA
ཇ	0F47	13AC 0020 0002	TIBETAN LETTER JA
མཇ	0F58 0F47	13AE 0020 0002	TIBETAN LETTER MA + TIBETAN LETTER JA
འཇ	0F60 0F47	13AF 0020 0002	TIBETAN LETTER –A + TIBETAN LETTER JA
རྗ	0F62 0F97	13B0 0020 0002	TIBETAN LETTER RA + TIBETAN SUBJOINED LETTER JA

ལྗ	0F63 0F97	13B1 0020 0002	TIBETAN LETTER LA +TIBETAN SUBJOINED LETTER JA
བརྗ	0F56 0F62 0F97	13B2 0020 0002	TIBETAN LETTER BA + TIBETAN LETTER RA + TIBETAN SUBJOINED LETTER JA
ཉ	0F49	13B3 0020 0002	TIBETAN LETTER NYA
གཉ	0F42 0F49	13B5 0020 0002	TIBETAN LETTER GA +TIBETAN LETTER NYA
མཉ	0F58 0F49	13B6 0020 0002	TIBETAN LETTER MA + TIBETAN LETTER NYA
རྙ	0F62 0F99	13B8 0020 0002	TIBETAN LETTER RA +TIBETAN SUBJOINED LETTER NYA
སྙ	0F66 0F99	13B9 0020 0002	TIBETAN LETTER SA +TIBETAN SUBJOINED LETTER NYA
བརྙ	0F56 0F62 0F99	13C0 0020 0002	TIBETAN LETTER BA +TIBETAN LETTER RA +TIBETAN SUBJOINED LETTER NYA
བསྙ	0F56 0F66 0F99	13C3 0020 0002	TIBETAN LETTER BA +TIBETAN LETTER SA +TIBETAN SUBJOINED LETTER NYA
ཏ	0F4F	13C6 0020 0002	TIBETAN LETTER TA
གཏ	0F42 0F4F	13C9 0020 0002	TIBETAN LETTER GA + TIBETAN LETTER TA
བཏ	0F56 0F4F	13CA 0020 0002	TIBETAN LETTER BA + TIBETAN LETTER TA
རྟ	0F62 0F9F	13CC 0020 0002	TIBETAN LETTER RA +TIBETAN SUBJOINED LETTER TA
ལྟ	0F63 0F9F	13D0 0020 0002	TIBETAN LETTER LA +TIBETAN SUBJOINED LETTER TA
སྟ	0F66 0F9F	13D3 0020 0002	TIBETAN LETTER SA +TIBETAN SUBJOINED LETTER TA
བརྟ	0F56 0F62 0F9F	13D6 0020 0002	TIBETAN LETTER BA +TIBETAN LETTER RA +TIBETAN SUBJOINED LETTER TA
བསྟ	0F56 0F66 0F9F	13D9 0020 0002	TIBETAN LETTER BA +TIBETAN LETTER SA +TIBETAN SUBJOINED LETTER TA
ཐ	0F50	13DC 0020 0002	TIBETAN LETTER THA
མཐ	0F58 0F50	13DF 0020 0002	TIBETAN LETTER MA +TIBETAN LETTER THA
འཐ	0F60 0F50	13E0 0020 0002	TIBETAN LETTER –A +TIBETAN LETTER THA
ད	0F51	13E2 0020 0002	TIBETAN LETTER DA
གད	0F42 0F51	13E4 0020 0002	TIBETAN LETTER GA +TIBETAN LETTER DA

བད	0F56 0F51	13E6 0020 0002	TIBETAN LETTER BA +TIBETAN LETTER DA
མད	0F58 0F51	13E8 0020 0002	TIBETAN LETTER MA +TIBETAN LETTER DA
འད	0F60 0F51	13EA 0020 0002	TIBETAN LETTER –A +TIBETAN LETTER DA
རྡ	0F62 0FA1	13EC 0020 0002	TIBETAN LETTER RA +TIBETAN SUBJOINED LETTER DA
ལྡ	0F63 0FA1	13EF 0020 0002	TIBETAN LETTER LA +TIBETAN SUBJOINED LETTER DA
སྡ	0F66 0FA1	13F0 0020 0002	TIBETAN LETTER SA +TIBETAN SUBJOINED LETTER DA
བརྡ	0F56 0F62 0FA1	13F2 0020 0002	TIBETAN LETTER BA +TIBETAN LETTER RA +TIBETAN SUBJOINED LETTER DA
བལྡ	0F56 0F63 0FA1	13F 4 0020 0002	TIBETAN LETTER BA +TIBETAN LETTER LA +TIBETAN SUBJOINED LETTER DA
བསྡ	0F56 0F66 0FA1	13F6 0020 0002	TIBETAN LETTER BA +TIBETAN LETTER SA +TIBETAN SUBJOINED LETTER DA
ན	0F53	13F8 0020 0002	TIBETAN LETTER NA
གན	0F42 0F53	13FA 0020 0002	TIBETAN LETTER GA +TIBETAN LETTER NA
མན	0F58 0F53	13FC 0020 0002	TIBETAN LETTER MA +TIBETAN LETTER NA
རྣ	0F62 0FA3	13FF 0020 0002	TIBETAN LETTER RA +TIBETAN SUBJOINED LETTER NA
སྣ	0F66 0FA3	1400 0020 0002	TIBETAN LETTER SA +TIBETAN SUBJOINED LETTER NA
བརྣ	0F56 0F62 0FA3	1402 0020 0002	TIBETAN LETTER BA +TIBETAN LETTER RA +TIBETAN SUBJOINED LETTER NA
བསྣ	0F56 0F66 0FA3	1404 0020 0002	TIBETAN LETTER BA +TIBETAN LETTER SA +TIBETAN SUBJOINED LETTER NA
པ	0F54	1406 0020 0002	TIBETAN LETTER PA
དཔ	0F51 0F54	1408 0020 0002	TIBETAN LETTER DA + TIBETAN LETTER PA
ལྤ	0F63 0FA4	140A 0020 0002	TIBETAN LETTER LA + TIBETAN SUBJOINED LETTER PA
སྤ	0F66 0FA4	140C 0020 0002	TIBETAN LETTER SA +TIBETAN SUBJOINED LETTER PA
ཕ	0F55	140F 0020 0002	TIBETAN LETTER PHA
འཕ	0F60 0F55	1410 0020 0002	TIBETAN LETTER –A + TIBETAN LETTER PHA

བ	0F56	1412 0020 0002	TIBETAN LETTER BA
དབ	0F51 0F56	1414 0020 0002	TIBETAN LETTER DA +TIBETAN LETTER BA
འབ	0F60 0F56	1416 0020 0002	TIBETAN LETTER –A + TIBETAN LETTER BA
རྦ	0F62 0FA6	1418 0020 0002	TIBETAN LETTER RA + TIBETAN SUBJOINED LETTER BA
ལྦ	0F63 0FA6	141A 0020 0002	TIBETAN LETTER LA +TIBETAN SUBJOINED LETTER BA
སྦ	0F66 0FA6	141C 0020 0002	TIBETAN LETTER SA + TIBETAN SUBJOINED LETTER BA
མ	0F58	141F 0020 0002	TIBETAN LETTER MA
དམ	0F51 0F58	1420 0020 0002	TIBETAN LETTER DA + TIBETAN LETTER MA
རྨ	0F62 0FA8	1422 0020 0002	TIBETAN LETTER RA +TIBETAN SUBJOINED LETTER MA
སྨ	0F66 0FA8	1424 0020 0002	TIBETAN LETTER SA + TIBETAN SUBJOINED LETTER MA
ཙ	0F59	1426 0020 0002	TIBETAN LETTER TSA
གཙ	0F42 0F59	1428 0020 0002	TIBETAN LETTER GA + TIBETAN LETTER TSA
བཙ	0F56 0F59	142A 0020 0002	TIBETAN LETTER BA +TIBETAN LETTER TSA
རྩ	0F62 0FA9	142C 0020 0002	TIBETAN LETTER RA +TIBETAN SUBJOINED LETTER TSA
སྩ	0F66 0FA9	142F 0020 0002	TIBETAN LETTER SA +TIBETAN SUBJOINED LETTER TSA
བརྩ	0F56 0F62 0FA9	1430 0020 0002	TIBETAN LETTER RA +TIBETAN SUBJOINED LETTER TSA
བསྩ	0F56 0F66 0FA9	1432 0020 0002	TIBETAN LETTER BA +TIBETAN LETTER SA +TIBETAN SUBJOINED LETTER TSA
ཚ	0F5A	1434 0020 0002	TIBETAN LETTER TSHA
མཚ	0F58 0F5A	1436 0020 0002	TIBETAN LETTER MA + TIBETAN LETTER TSHA
འཚ	0F60 0F5A	1438 0020 0002	TIBETAN LETTER –A + TIBETAN LETTER TSHA
ཛ	0F5B	143A 0020 0002	TIBETAN LETTER DZA
མཛ	0F58 0F5B	143C 0020 0002	TIBETAN LETTER MA + TIBETAN LETTER DZA

འཛ	0F60 0F5B	143F 0020 0002	TIBETAN LETTER –A +TIBETAN LETTER DZA
རྫ	0F62 0FAB	1440 0020 0002	TIBETAN LETTER RA +TIBETAN SUBJOINED LETTER DZA
བརྫ	0F56 0F62 0FAB	1442 0020 0002	TIBETAN LETTER BA +TIBETAN SUBJOINED LETTER DZA
ཝ	0F5D	1444 0020 0002	TIBETAN LETTER WA
ཞ	0F5E	1446 0020 0002	TIBETAN LETTER ZHA
གཞ	0F42 0F5E	1448 0020 0002	TIBETAN LETTER GA +TIBETAN LETTER ZHA
བཞ	0F56 0F5E	144A 0020 0002	TIBETAN LETTER BA +TIBETAN LETTER ZHA
ཟ	0F5F	144C 0020 0002	TIBETAN LETTER ZA
གཟ	0F42 0F5F	144F 0020 0002	TIBETAN LETTER GA +TIBETAN LETTER ZA
བཟ	0F56 0F5F	1450 0020 0002	TIBETAN LETTER BA +TIBETAN LETTER ZA
འ	0F60	1452 0020 0002	TIBETAN LETTER –A
ཡ	0F61	1454 0020 0002	TIBETAN LETTER YA
གཡ	0F42 0F61	1456 0020 0002	TIBETAN LETTER GA +TIBETAN LETTER YA
ར	0F62	1458 0020 0002	TIBETAN LETTER RA
ཪ	0F6A	145A 0020 0002	TIBETAN LETTER FIXED-FORM RA
བཪ	0F56 0F6A	145C 0020 0002	TIBETAN LETTER BA +TIBETAN LETTER FIXED-FORM RA
ལ	0F63	145A 0020 0002	TIBETAN LETTER LA
ཤ	0F64	1460 0020 0002	TIBETAN LETTER SHA
གཤ	0F42 0F64	1462 0020 0002	TIBETAN LETTER GA +TIBETAN LETTER SHA
བཤ	0F56 0F64	1464 0020 0002	TIBETAN LETTER BA +TIBETAN LETTER SHA
ས	0F66	1466 0020 0002	TIBETAN LETTER SA
གས	0F42 0F66	1468 0020 0002	TIBETAN LETTER GA +TIBETAN LETTER SA

བས	0F56 0F66	146A 0020 0002	TIBETAN LETTER BA +TIBETAN LETTER SA
ཧ	0F67	146C 0020 0002	TIBETAN LETTER HA
ལྷ	0F63 0FB7	146F 0020 0002	TIBETAN LETTER LA + TIBETAN SUBJOINED LETTER HA
ཨ	0F68	1470 0020 0002	TIBETAN LETTER A
ཊ	0F4A	13C6 0020 0008	TIBETAN LETTER TTA
ཋ	0F4B	13DC 0020 0008	TIBETAN LETTER TTHA
ཌ	0F4C	13E2 0020 0008	TIBETAN LETTER DDA
ཎ	0F4E	13F8 0020 0008	TIBETAN LETTER NNA
◌ཾ	0F7E	147A 0020 0002	TIBETAN SIGN RJES SU NGA RO
◌ྂ	0F82	147C 0020 0002	TIBETAN SIGN NYI ZLA NAA DA
◌ྃ	0F83	147F 0020 0002	TIBETAN SIGN SNA LDAN
ཥ	0F65	1480 0020 0002	TIBETAN LETTER SSA
ཀྵ	0F69	1482 0020 0002	TIBETAN LETTER KSSA
◌ི	0F72	1484 0020 0002	TIBETAN VOWEL SIGN I
◌ྀ	0F80	1486 0020 0002	TIBETAN VOWEL SIGN REVERSED I
◌ུ	0F74	1488 0020 0002	TIBETAN VOWEL SIGN U
◌ེ	0F7A	148A 0020 0002	TIBETAN VOWEL SIGN E
◌ཻ	0F7B	148C 0020 0002	TIBETAN VOWEL SIGN EE
◌ོ	0F7C	148F 0020 0002	TIBETAN VOWEL SIGN O
◌ཽ	0F7D	1490 0020 0002	TIBETAN VOWEL SIGN OO
◌ྐ	0F90	1492 0020 0002	TIBETAN SUBJOINED LETTER KA
◌ྑ	0F91	1494 0020 0002	TIBETAN SUBJOINED LETTER KHA

◌ྒ	0F92	1496 0020 0002	TIBETAN SUBJOINED LETTER GA
◌ྔ	0F94	1498 0020 0002	TIBETAN SUBJOINED LETTER NGA
◌ྕ	0F95	149A 0020 0002	TIBETAN SUBJOINED LETTER CA
◌ྖ	0F96	149C 0020 0002	TIBETAN SUBJOINED LETTER CHA
◌ྗ	0F97	149F 0020 0002	TIBETAN SUBJOINED LETTER JA
◌ྙ	0F99	1500 0020 0002	TIBETAN SUBJOINED LETTER NYA
◌ྟ	0F9F	1502 0020 0002	TIBETAN SUBJOINED LETTER TA
◌ྛ	0F9B	1505 0020 0002	TIBETAN SUBJOINED LETTER TTHA
◌ྡ	0FA1	1506 0020 0002	TIBETAN SUBJOINED LETTER DA
◌ྣ	0FA3	1508 0020 0002	TIBETAN SUBJOINED LETTER NA
◌ྤ	0FA4	150A 0020 0002	TIBETAN SUBJOINED LETTER PA
◌ྥ	0FA5	150C 0020 0002	TIBETAN SUBJOINED LETTER PHA
◌ྦ	0FA6	150F 0020 0002	TIBETAN SUBJOINED LETTER BA
◌ྨ	0FA8	1510 0020 0002	TIBETAN SUBJOINED LETTER MA
◌ྩ	0FA9	1512 0020 0002	TIBETAN SUBJOINED LETTER TSA
◌ྪ	0FAA	1514 0020 0002	TIBETAN SUBJOINED LETTER TSHA
◌ྫ	0FAB	1516 0020 0002	TIBETAN SUBJOINED LETTER DZA
◌ྭ	0FAD	1518 0020 0002	TIBETAN SUBJOINED LETTER WA
◌ྮ	0FAE	151A 0020 0002	TIBETAN SUBJOINED LETTER ZHA
◌ྯ	0FAF	151C 0020 0002	TIBETAN SUBJOINED LETTER ZA
◌ྰ	0FB0	151F 0020 0002	TIBETAN SUBJOINED LETTER –A
◌ྱ	0FB1	1520 0020 0002	TIBETAN SUBJOINED LETTER YA

◌ྲ	0FB2	1522 0020 0002	TIBETAN SUBJOINED LETTER RA
◌ླ	0FB3	1524 0020 0002	TIBETAN SUBJOINED LETTER LA
◌ྴ	0FB4	1526 0020 0002	TIBETAN SUBJOINED LETTER SHA
◌ྶ	0FB6	1528 0020 0002	TIBETAN SUBJOINED LETTER SA
◌ྷ	0FB7	152A 0020 0002	TIBETAN SUBJOINED LETTER HA
◌ྸ	0FB8	152C 0020 0002	TIBETAN SUBJOINED LETTER A
◌ྚ	0F9A	152F 0020 0002	TIBETAN SUBJOINED LETTER TTA
◌ྛ	0F9B	1530 0020 0002	TIBETAN SUBJOINED LETTER TTHA
◌ྜ	0F9C	1532 0020 0002	TIBETAN SUBJOINED LETTER DDA
◌ྞ	0F9E	1534 0020 0002	TIBETAN SUBJOINED LETTER NNA
◌ྵ	0FB5	1536 0020 0002	TIBETAN SUBJOINED LETTER SSA
◌ྐྵ	0F90 0FB5	1538 0020 0002	TIBETAN SUBJOINED LETTER
।	0F0D	1371 0020 0002	TIBETAN MARK SHAD
༎	0F0E	1372 0020 0002	TIBETAN MARK NYIS SHAD
◌྄	0F84	0000 00C4 0002	TIBETAN MARK HALANTA
◌ཱ	0F71	0000 00C9 0002	TIBETAN VOWEL SIGN AA
◌༹	0F39	0000 00CA 0002	TIBETAN MARK TSA-PHRU
◌ཿ	0F7F	0000 00CB 0002	TIBETAN SIGN RNAM BCAD
྅	0F85	0000 00CD 0002	TIBETAN MARK PALUTA
ྈ	0F88	0000 00D5 0002	TIBETAN SIGN LCE TSA CAN
ྉ	0F89	0000 00D8 0002	TIBETAN SIGN MACHU CAN
ྊ	0F8A	0000 00DA 0002	TIBETAN SIGN GURU CAN RGYINGS

ྋ	0F8B	0DC7 0020 0002	TIBETAN SIGN GURU MED RGYINGS
༠	0F20	0DC8 0020 0002	TIBETAN DIGIT ZERO
༳	0F33	0350 0020 0002	TIBETAN DIGIT HALF ZERO
༡	0F21	0E12 0020 0002	TIBETAN DIGIT ONE
༪	0F2A	0E13 0020 0002	TIBETAN DIGIT HALF ONE
༢	0F22	0E29 0020 0002	TIBETAN DIGIT TWO
༫	0F2B	0E2A 0020 0002	TIBETAN DIGIT HALF TWO
༣	0F23	0E2A 0020 0002	TIBETAN DIGIT THREE
༬	0F2C	0E2B 0020 0002	TIBETAN DIGIT HALF THREE
༤	0F24	0E2B 0020 0002	TIBETAN DIGIT FOUR
༭	0F2D	0E2C 0020 0002	TIBETAN DIGIT HALF FOUR
༥	0F25	0E2C 0020 0002	TIBETAN DIGIT FIVE
༮	0F2E	0E2D 0020 0002	TIBETAN DIGIT SIX
༦	0F26	0E2D 0020 0002	TIBETAN DIGIT HALF SIX
༯	0F2F	0E2E 0020 0002	TIBETAN DIGIT SEVEN
༧	0F27	0E2F 0020 0002	TIBETAN DIGIT HALF SEVEN
༰	0F30	0E30 0020 0002	TIBETAN DIGIT HALF SEVEN
༨	0F28	0E31 0020 0002	TIBETAN DIGIT EIGHT
༱	0F31	0E32 0020 0002	TIBETAN DIGIT HALF EIGHT
༩	0F29	0E33 0020 0002	TIBETAN DIGIT NINE
༲	0F32	0E34 0020 0002	TIBETAN DIGIT HALF NINE
་	0F0B	1370 0020 0002	TIBETAN MARK INTERSYLLABIC TSHEG

NB	0F0C	1373 0020 0002	TIBETAN MARK DELIMETER TSHEG BSTAR
ཀྵ	0F69	[1375 0020 0002], [1536 0020 0002]	TIBETAN LETTER KSSA
◌ཱི	0F73	[0000 00C9 0002], [1484 0020 0002]	TIBETAN VOWEL SIGN II
◌ྡྷ	0FA2	[1506 0020 0002], [152A 0020 0002]	TIBETAN SUB-JOINED LETTER DHA
◌ྦྷ	0FA7	[150F 0020 0002], [152A 0020 0002]	TIBETAN SUB-JOINED LETTER BHA
◌ླྀ	0F78	[1524 0020 0002], [1486 0020 0002]	TIBETAN VOWEL SIGN VOCALIC L
◌ྲཱྀ	0F77	[1522 0020 0002], [0000 00C9 0002], [1486 0020 0002]	TIBETAN VOWEL SIGN VOCALIC RR
◌ཱུ	0F75	[0000 00C9 0002], [1488 0020 0002]	TIBETAN VOWEL SIGN UU
◌ྲྀ	0F76	[1522 0020 0002], [151F 0020 0002]	TIBETAN VOWEL SIGN VOCALIC R
◌ླཱྀ	0F79	[1524 0020 0002], [0000 00C9 0002], [1486 0020 0002]	TIBETAN VOWEL SIGN VOCALIC LL
◌ཱྀ	0F81	[0000 00C9 0002], [1486 0020 0002]	TIBETAN VOWEL SIGN REVERSED II
◌ྒྷ	0F93	[1496 0020 0002], [152A 0020 0002]	TIBETAN SUB-JOINED LETTER GHA
◌ྫྷ	0FAC	[1516 0020 0002], [152A 0020 0002]	TIBETAN SUBJOINED LETTER DZHA
◌ྜྷ	0F9D	[1532 0020 0002], [152A 0020 0002]	TIBETAN SUBJOINED LETTER DDHA
◌ྐྵ	0FB9	[1492 0020 0002], [1536 0020 0002]	TIBETAN SUBJOINED LETTER KSSA
ཛྷ	0F5C	[143A 0020 0002], [152A 0020 0002]	TIBETAN LETTER DZHA
བྷ	0F57	[1412 0020 0002], [152A 0020 0002]	TIBETAN LETTER BHA
དྷ	0F52	[13E2 0020 0002], [152A 0020 0002]	TIBETAN LETTER DHA
ཌྷ	0F4D	[1476 0020 0002], [152A 0020 0002]	TIBETAN LETTER DDHA
གྷ	0F43	[1386 0020 0002], [152A 0020 0002]	TIBETAN LETTER GHA

4.2.3. Results

This section shows some results obtained by sorting a sample input based on the collation elements defined.

Table 4.9. Input and Corresponding Sorted Output for Dzongkha

Input		Output	
རྐུན་ཇག	བཀུགཔ།	ཀ་ཀུ་ཙུ	དཀར་ཁུང་།
ཀི་སྒྲ།	སྐྱུ་སྒོགཔ།	ཀ་རྐྱང་།	དཀར་ཁྲ།
བརྐོས་སྤྱིར།	བཀོལ་བདེ།	ཀ་ཙུང་།	དཀོན་གཉེར།
ཀ་ཀུ་ཙུ	ཀྲང་ཀྲིང་།	ཀ་ཙུང་ནང་།	དཀོན་ཐག་ཐ།
ཀཱུ་ཀོ་ལ།	དཀར་ཁ།	ཀ་ཙོག་ཞང་གསུམ།	དཀྱིལ་ཆོག
བསྐྱེད་རིམ།	ཀྲེབ་ཀྲེམ།	ཀི་སྒྲ།	དཀྱུས།
ལྐོག་ཆད།	ཀྲོ་དོམ།	ཀི་ཙུ་རམ།	དཀྱེལ།
ཀི་ཙུ་རམ།	ཀླུ་སྒྲུབ།	ཀུན་དགའ་རྒྱལ་མཚན།	དཀྲི་ཤིང་།
ཀླུ་སྒྲུབ་སྙིང་པ།	ཀྱུག་ཀྱོག	ཀུན་དགའ་རྒྱལ་མོ།	བཀག་ཀ།
ཀ་རྐྱང་།	དཀར་སྦྲ།	ཀོ་སྤྱིན།	བཀུགན།
དཀོན་ཐག་ཐ།	ཀློང་བརྡུར།	ཀོ་ཕི།	བཀུགཔ།
ཀླ་ཀློའི་སྐད།	དཀོན་གཉེར།	ཀྭ་ལས།	བཀོལ་བདེ།
བཀྲོངས།	སྐུ་མཚམས།	ཀཱུ་ཀོ་ལ།	བཀྱིགས།
ཀྱུ།	དཀྱུས།	ཀཱུ་ལི།	བཀྲ་ཤིས།
ཀླ་ཀློཔ།	དཀྱེལ།	ཀྱུག་ཀྱོག	བཀྱེ་བ།
ཀྭ་ལས།	ཀ་ཙུང་ནང་།	ཀྱི་ལི་ལི།	བཀྱོན།
ཀུན་དགའ་རྒྱལ་མོ།	བཀག་ཀ།	ཀྱི་ཧུད།	བཀྲག
དཀར་ཁྲ།	བཀྱིགས།	ཀྱུ།	བཀྲོངས།

ཀོ་ཁི།	བཀྲ་ཤིས།	ཀྱེ་རྡོ་རྗེ།	རྐ་ཆུ།
ཀ་ཙོག་ཞང་གསུམ།	བཀྱེ་བ།	ཀྲང་ཀྲི།	རྐུན་ཇག
ཀ་ཙུང་།	བཀྱོན།	ཀྲང་ཀྲིང་།	རྐུན།
ཀྲང་ཀྲི།	བཀྲག	ཀྲང་ཀྲོང་།	ལྐུག་རྟགས།
དཀར་ཁུང་།	རྐ་ཆུ།	ཀྲི་མི།	ལྐོག་ཆད།
བཀུགན།	རྐུན།	ཀྲ ུ་སྐུད་ལི།	སྐད་ཅན་པ།
ཀཱ་ལི།	ལྐུག་རྟགས།	ཀྲེབ་ཀྲེམ།	སྐད་གཉིས་པ།
ཀོ་སྐྲིན།	ཀླུ་རྒྱལ།	ཀྲོ་རྡོམ།	སྐུ་མཚམས།
སྐད་གཉིས་པ།	སྐད་ཅན་པ།	ཀླ་ཀློཔ།	སྐེ་ཤིང་།
ཀྱི་ལི་ལི།	སྐེ་ཤིང་།	ཀླ་ཀློའི་སྐད།	སྐོར་ལས།
དཀྱིལ་ཆོག	སྐོར་ལས།	ཀླུ་རྒྱལ།	སྐྱུ་སྐོ་གཔ།
ཀྱི་ཧུད།	སྐྱིད་སྡུག	ཀླུ་སྒྲུབ།	སྐྱིད་སྡུག
ཀྱེ་རྡོ་རྗེ།	སྐྱུ་རུ་ར།	ཀླུ་སྒྲུབ་སྙིང་པ།	སྐྱུ་རུ་ར།
དཀྲི་ཤིང་།	ཀུན་དགའ་རྒྱལ་མཚན།	ཀློང་རྒྱུར།	བསྐོས་སྟེར།
ཀྲ ུ་སྐུད་ལི།	ཀྲང་ཀྲོང་།	དཀར་སྦྱ།	བསྐྱེད་རིམ།
ཀྲི་མི།		དཀར་ཁ།	

4.3. Conclusion

Dzongkha resembles with both Indic and South East Asian languages. Like Indic languages sorting is mainly carried out at primary level. Like Lao, collation in Dzongkha is based on intricate syllable structure. Collation element assignment requires context sensitive analysis. Numerals and their half forms are given lightest weights. These are followed by consonants and their respective conjuncts, then vowels followed by sub-joined forms of consonants. Few special signs are given secondary weights. The sort order has been based on Dzongkha Dictionary from Dzongkha Development Authority [34].

5. Lao

Lao language is derived from Kam-Tai branch of Tai-Kadai language spoken by approximately 3 million people in Laos and Thailand [17]. Traditional Lao literature has been written in Lao and Tham scripts. The Lao script emerged in 13th [18] or 14th century [19], deriving mutually with old Thai script from Brahmi writing system. Lao script was simplified in 1960, making it more regular [18].

5.1. *Writing System*

5.1.1. Character Set

Like Indic scripts, Lao script consonants also carry an inherent vowel, and in addition an inherent tone, both of which can be over-ridden by explicitly specifying them. Lao script has 27 consonants which are divided into three classes, high, middle and low. This grouping helps in determining the tone of the syllable, along with the tone marks and vowels. These consonants are given in Figure 5.1. Vowels are always written around a central consonant. Vowels occur in full form or as marks which can attach before, after, above or below the consonant. Lao vowels are shown in Figure 5.2 [21]. Slightly variant vowel list is reported in [4].

ກ ຂ ຄ ງ ຈ ສ ຊ ຍ ດ ຕ ຖ ທ ນ ບ ປ ຜ ຝ ພ ຟ ມ ຢ ຣ ລ ວ ຫ ອ ຮ

Figure 5.1. Lao Consonants

ະ ◌ິ ◌ຶ ◌ຸ ເXະ ແXະ ໂXະ ເXາະ ເ ◌ິ ເ ◌ັຍ ເ◌ຶອ ◌ົວະ

Short Vowels

Xາ ◌ີ ◌ື ◌ູ ເX ແX ໂX ◌ໍ ເ ◌ີ ເXຍ ເ ◌ືອ ◌ົວ

Long Vowels

ໄX ໃX ເ ◌ົາ ◌ໍາ

Diphthongs

Figure 5.2. Lao Vowels [21]
(X used as a placeholder for a consonant)

Lao script also has four tone marks, shown in Figure 5.3.

◌່ ◌້ ◌໊ ◌໋

Figure 5.3. Lao Tone Marks

Lao also possesses special characters shown in Table 5.1.

Table 5.1. Lao Special Characters

Name	Glyph
Mai Sum (Sentence Repetition)	ຯ
Mai Sum (Word Repetition)	ໆ
Mai Kalan	◌໌

Mai Sum (ຯ, ໆ) are used for sentence and word repetition. These are used instead of writing the whole sentence or whole word again. Mai Kalan is used with foreign words and is optional.

Lao has its own set of numerals given in Figure 5.4.

໐ ໑ ໒ ໓ ໔ ໕ ໖ ໗ ໘ ໙

Figure 5.4. Lao Digits

5.1.2. Script Details

5.1.2.1. No Word Spacing

Like other South-East Asian scripts such as Thai and Burmese, Lao does not have spaces between words. Native readers identify word boundaries using their tacit knowledge of the language. Text is written in continuum and space is only used at the end of sentence or clause.

5.1.2.2. Vowel and Tone Marks

Vowels are used in conjunction with consonants to modify the way they are pronounced. They attach at the front, back, top or bottom of the consonant. Unlike Indic languages multiple vowels can attach to a consonant. These variations are shown in Table 5.2.

Table 5.2. Lao Vowels with Consonant KO

ເ+ກ	ເກ	Connects to Left
ກ+ະ	ກະ	Connects to Right
ກ+◌ຸ	ກຸ	Connects at Bottom
ກ+◌ິ	ກິ	Connects at Top
ເ+ກ+ ◌ີ	ເກີ	Connects to Left and Top

The tone marks are always placed above the consonants. If there is already a vowel above consonant, the tone mark will stack above the vowel, as shown in Table 5.3.

Table 5.3. Placement of Lao Tone Marks

ເ+ກ+◌່	ເກ່	Above the Consonant
ກ+◌ິ+◌້	ກິ້	Above the Vowel

Further details are given in the discussion on syllable structure later.

5.1.2.3. Syllable and Syllabification

Lao is a syllable based language. The syllables are structured around a central consonant (also known as main or nuclear consonant). A syllable might optionally have combinational consonants, at least one vowel which may be placed before, after, above or below the main consonant, and up to one tone mark. This is illustrated in Figure 5.5 below. Capital C indicates the nuclear consonant. The subscripts “0..n” mean zero to *n*, indicating that all are optional (in case of zero) except the nuclear ***C***.

		$T_{0,1}$		
		$V_{0,1}$		
$V_{0,1}$	$C_{0,1}$	***C***	$V_{0,1}$	$C_{0..4}$
		$C_{0,1}$		
		$V_{0,1}$		

Figure 5.5. Generic Syllable Structure for Lao (C = Consonant; V = Vowel; T = Tone Mark)

A detailed syllable template for Lao is shown Figure 5.6. X_0 through X_{10} are explained below.

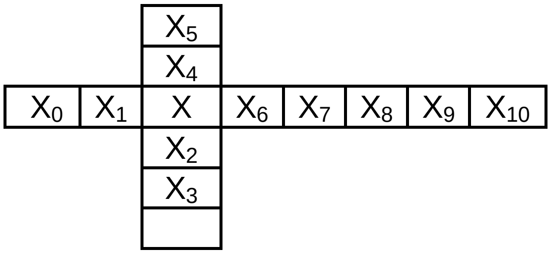

Figure 5.6. Detailed Syllable Structure for Lao

- X_0 represents a vowel which always occurs before the nuclear consonant X.
- X1 is a combination consonant ຫ which comes before the nuclear consonant, only if the nuclear consonant is one of {ງ, ຍ, ນ, ມ, ລ, ວ}. It can also occur before ◌ຼ.
- X represents the nuclear consonants.
- X2 is ◌ຼ and comes only when ຫ occurs as X1 (in this case, there will be no nuclear consonant) and the combination forms the nuclear consonant.
- X_3 represents vowels which occur under the nuclear consonant.
- X_4 represents vowel which occur above the nuclear consonant.
- X_5 represents tone marks which appear above nuclear consonant or above vowels.
- X_6 represents consonant vowel, which occurs after nuclear consonant. This functions as vowel when the syllable does not have any vowels, and always appear with X_8.
- X_7 represents an after-vowel. However X_{71} always indicates the end of syllable and it never exists with a tone mark.
- X_8 represents alternate consonants.
- X_9 represents alternate consonant to pronounce foreign language words. It always exists with X_{10}.

- X_{10} represents different marks as discussed in Table 5.1. Mai Sum may be considered outside the syllable.

The following Table 5.4 further classifies where each Lao character can occur. A character can fall under multiple categories depending upon its position in syllable.

Table 5.4. Positional Restrictions on Lao Characters in a Syllable

X_0	X_1	X	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
ເ X_{01} ແ X_{02} ໂ X_{03} ໄ X_{04} ໃ X_{05}	ຫ	ກ ຂ ຄ ງ ຈ ຊ ຍ ດ ຕ ຖ ບ ປ ຜ ຝ ພ ຟ ມ ຢ ຣ ລ ຫ ອ ຮ ໜ ໝ ວ ສ ທ ນ	◌ຼ	◌ຸ ◌ູ	◌ິ X_{41} ◌ີ X_{42} ◌ຶ X_{43} ◌ື X_{44} ◌ໍ X_{45} ◌ົ X_{46} ◌ັ X_{47}	◌່ ◌້ ◌໊ ◌໋	ວ X_{61} ອ X_{62} ຽ X_{63}	ະ X_{71} າ X_{72} ◌ຳ X_{73}	ກ ງ ຍ ດ ນ ມ ບ ວ	ຈ ສ ຊ ພ ຟ ລ	ໆ ຯ ◌໌

Syllable boundaries are detected based on a set of conditions. For example the syllable ເກິ່ດ satisfies condition: $X_{01}(X_1)$ $X(X_2)$ X_{4_1} | X_{4_2} (X_5) (X_8) $(X_9$: $X_{103})$ $(X_{10\_1}$ | $X_{10\_2})$. It states that a syllable that fulfills this condition must have vowel X_{01} ເ. Combinational consonants X_1 and X_2 are optional. It should have a main consonant X which is ກ in this example string. It must have one of the two vowels X_{41} or X_{42} (◌ິ or ◌ີ). Tone mark X_5 and consonants X_8 and X_9 are also optional. Moreover if X_9 occurs it must be followed by $X_{103.}$ One of the X_{101} or X_{102} can occur optionally. The syllable template is filled for this string in Figure 5.7.

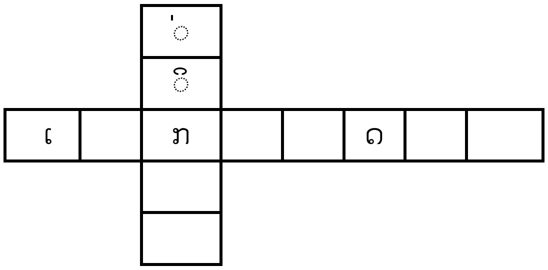

Figure 5.7. Syllable Template Filled for Lao String ເກິ່ດ

Further algorithmic details and a complete set of syllabification rules are given in [20].

5.2. Collation

Two different strategies are commonly used in Lao for collation. One of these uses base characters and collapses them into bigger linguistic units and assigns a single collation element per unit. The second strategy does not collapse the input characters and assigns a single collation element to each character in the script. The two mechanisms are known as language based versus script based collation.

Lao language has syllable based collation. The word is subdivided into a sequence of syllables for sorting. Then, given two words, their initial syllables are compared. The second syllables of these words are only compared if the first syllables are identical, and so on. This strategy is significantly different from Unicode Collation Algorithm [2] discussed in Chapter 2. In the earlier algorithm, after collation elements are assigned, a single sort key is generated for each word for a single comparison with other sort keys from other words. However, in the case of Lao, there will be a sort key generated for each syllable (not word!). The comparison of words will be an iterative process which compares sort keys of each syllable in one word in sequence, with corresponding sort keys of syllables in other word. These comparisons will be done until a difference is found.

Within the syllable, Lao sorts at four levels, with nuclear consonants getting the primary weight, vowels getting the secondary weight, alternate consonants getting the tertiary weight and tone marks getting the quaternary level weight. Punctuation marks and some other Lao characters are ignorable at all levels. Most popular Lao dictionaries generally agree on the order of consonants, though differences lie in the ordering of vowels and combining consonants.

5.2.1. Text Processing

5.2.1.1. Syllabification

Lao strings are collated based on syllable sequence. Thus, it is critical to syllabify the strings to be compared. This can be done through advanced language processing techniques, both by extensive rule-based systems [4], or using statistical methods. Some initial details for rule based syllabification are provided in Section 5.1.2.3. There has been very limited work on statistical solutions for Lao syllabification.

5.2.1.2. Syllable Parsing

Lao characters behave differently in collation depending on where they occur in a syllable. For example, consonants get primary weight in collation if they occur as main consonant X, combinational consonant X_1, X2 but tertiary weight if they occur in a secondary role as X_8 and X_9, as given in Table 5.4. Thus, the syllabification process should not only return syllable boundaries but also label the role of each character within the syllable string. This implies that complete internal parsing of syllable is also desired. Lao Letter WO ວ can play as X, X_2, X_6 and X_8 in a syllable. Therefore it can acquire primary (when X or X_2), secondary (when X_6) or tertiary (when X_8) weight.

5.2.1.3. Reordering

The syllable structure illustrated in Figures 5.5 and 5.6 shows that a main consonant can be preceded optionally by another consonant and a dependent vowel. Like other Indic scripts, these characters are logically treated to occur after the central consonant. However, unlike encoding of South Asian scripts, Unicode encodes Lao characters in visual order (for backward compatibility with earlier systems for Thai and Lao). Thus, the characters have to be reordered into the logical order for collation.

In addition to typing the initial vowel and combinational consonant before the main consonant, there can be many ways of typing the characters following this main consonant in a syllable. For example, the string ຫູໍ can be generated by the sequence ຫ + ◌ູ + ◌ໍ or alternatively by the sequence ຫ + ◌ໍ + ◌ູ. These differences can also cause inconsistent collation results. This inconsistency in collation is shown in Table 5.5. below. Different character sequences result in different sort keys using the same collation elements causing different sorting order. Thus, reordering of all characters in a syllable needs to be conducted in a consistent order before sorting can proceed.

Table 5.5. Differences in Sort Keys Caused by Variation in Character Sequence

Syllable ເກ	Collation Elements	Sort Key
ກ+ເ	[0820 0200 0020 0002] [0000 021A 0020 0002]	[0820 0000 **0200 021A** 0000 0020 0020 0000 0002 0002]
ເ+ກ	[0000 021A 0020 0002] [0820 0200 0020 0002]	[0820 0000 **021A 0200** 0000 0020 0020 0000 0002 0002]

The desired order of characters in a syllable in Figure 5.6. is X (main consonant) X_1 X_2 (combinational consonants), X_0 X_3 X_4 X_6 X_7 (vowels), X_8 X_9 (alternate consonants), X_5 (tone marks) and X_{10} (special characters). As an example string ເນັ້ນ is reordered as ນ + ເ + ັ + ນ + ້ (X X_0 X_4 X_8 X_5). This reordering is performed after syllabification.

5.2.1.4. *Normalization*

A few characters inLao have multiple representations with Unicode encoding and thus normalization is required. The normalization is given in Table 5.6.

Table 5.6. Normalization in Lao

Decomposed Form	Unicodes of Decomposed Form	Equivalent Composed Form	Unicode of Composed Form
ຫ ນ	0EAB 0E99	ໜ	0EDC
ຫ ມ	0EAB 0EA1	ໝ	0EDD
◌ໍ າ	0ECD 0EB2	◌ຳ	0EB3

5.2.1.5. *Contraction of Consonants*

Lao letter Ho Sung ຫ combines with the consonants {ງ, ຍ, ນ, ມ, ຼ, ລ, ວ} to form different consonants which have their own collation weight. So these combinations undergo contraction and map onto a single collation element which is different from their individual collation elements. Lao letter ຼ is also grouped with these nuclear consonants. This is because it combines with ຫ in Lao words and is assigned the same collation element as ລ because ຫຼ is same as ຫລ.

5.2.1.6. *Contraction of Vowels*

Lao vowels are not encoded as shown in Figure 5.2. They are encoded in individual pieces, shown in Figure 5.8 below, for reasons of backward compatibility with existing systems of Lao and Thai. Thus, multiple encoded forms need to be combined together to form the Lao vowels.

ະ ◌ັ າ ◌ໍາ ◌ິ ◌ີ ◌ຶ ◌ື ◌ຸ ◌ູ ◌ົ ເ ແ ໂ ໃ ໄ

Figure 5.8. Encoded Characters and Marks for Forming Lao Vowels

However, each combined form maps onto a single vowel and thus a single collation element. Therefore, the contractions in Table 5.7 are needed to achieve Lao collation.

Table 5.7. Contraction to Single Collation Element from Multiple Encoded Characters

Glyph	Unicodes for Contraction
ເ + ◌ິ = ເ◌ິ	0EC0 + 0EB4
ເ + ◌ີ = ເ◌ີ	0EC0 + 0EB5
◌ົ + ວ + ະ ◌ົວະ	0EBB + 0EA7 + 0EB0
◌ົ + ວ ◌ົວ	0EBB + 0EA7
ເ + ◌ຶ + ອ = ເ◌ຶອ	0EC0 + 0EB6 + 0EAD
ເ + ◌ື + ອ = ເ◌ືອ	0EC0 + 0EB7 + 0EAD
ເ + ◌ົ + າ = ເ◌ົາ	0EC0 + 0EBB + 0EB2
ເ + ະ = ເXະ	0EC0 + 0EB0
ເ + າ + ະ = ເXາະ	0EC0 + 0EB2 + 0EB0
ແ + ະ = ແXະ	0EC1 + 0EB0
ໂ + ະ = ໂXະ	0EC2 + 0EB0
ເ + ຍ = ເXຍ	0EC0 + 0E8D

ເ + ◌ັ + ຍ ເ◌ັຍ	0EC0 + 0EB1 + 0E8D
◌ໍ + າ = ◌ໍາ	0ECD + 0EB2

5.2.2. Unicode Collation Elements

Lao language dictionaries follow two different collation sequences, which may be termed as Lao language-based (e.g. [22]) and script-based collation. Language-based collation uses the encoded vocalic symbols (given in Figure 5.8) to do the context based contractions (given in Table 5.7) to form singular vowels (given in Figure 5.2). A Collation element is then assigned to each vowel in Figure 5.2.

Script-based collation does not perform the contractions discussed but assigns collation element to each script symbol given in Figure 5.8. Thus, the collation is not done on basis of vowels but individual script symbols used for forming these vowels.

Syllabification, syllable based parsing, re-ordering and normalization is done in the same manner as discussed for both strategies. The difference in the strategies is just in contraction and eventual collation assignment process. Collation elements for the two strategies are also different and are given in Tables 5.8 and 5.9.

5.2.2.1. Language Based Sorting

The collation elements for language based sorting are given in Table 5.8.

Table 5.8. Lao Collation Elements for Language Based Sorting

Glyph	Unicode	Collation Elements	Unicode Name
← Consonants→			
ກ	0E81	0820 0200 0020 0002	LAO LETTER KO
ຂ	0E82	0822 0200 0020 0002	LAO LETTER KHO SUNG
ຄ	0E84	0824 0200 0020 0002	LAO LETTER KHO TAM
ງ	0E87	0826 0200 0020 0002	LAO LETTER NGO
ຈ	0E88	0828 0200 0020 0002	LAO LETTER CO

ສ	0EAA	082A 0200 0020 0002	LAO LETTER SO SUNG
ຊ	0E8A	082C 0200 0020 0002	LAO LETTER SO TAM
ຍ	0E8D	082E 0200 0020 0002	LAO LETTER NYO
ດ	0E94	0830 0200 0020 0002	LAO LETTER DO
ຕ	0E95	0832 0200 0020 0002	LAO LETTER TO
ຖ	0E96	0834 0200 0020 0002	LAO LETTER THO SUNG
ທ	0E97	0836 0200 0020 0002	LAO LETTER THO TAM
ນ	0E99	0838 0200 0020 0002	LAO LETTER NO
ບ	0E9A	083A 0200 0020 0002	LAO LETTER BO
ປ	0E9B	083C 0200 0020 0002	LAO LETTER PO
ຜ	0E9C	083E 0200 0020 0002	LAO LETTER PHO SUNG
ຝ	0E9D	0840 0200 0020 0002	LAO LETTER FO TAM
ພ	0E9E	0842 0200 0020 0002	LAO LETTER PHO TAM
ຟ	0E9F	0844 0200 0020 0002	LAO LETTER FO SUNG
ມ	0EA1	0846 0200 0020 0002	LAO LETTER MO
ຢ	0EA2	0848 0200 0020 0002	LAO LETTER YO
ຣ	0EA3	084A 0200 0020 0002	LAO LETTER LO LING
◌ຼ	0EBC	084C 0200 0020 0002	LAO SEMI VOWEL SIGN LO
ລ	0EA5	084E 0200 0020 0002	LAO LETTER LO LOOT
ວ	0EA7	0850 0200 0020 0002	LAO LETTER WO
ຫ	0EAB	0852 0200 0020 0002	LAO LETTER HO SUNG
ຫງ	0EAB+0E87	0854 0200 0020 0002	LAO LETTER HO SUNG+ LAO LETTER NGO
ຫຍ	0EAB+0E8D	0856 0200 0020 0002	LAO LETTER HO SUNG + LAO LETTER NYO
ຫນ	0EAB+0E99	0858 0200 0020 0002	LAO LETTER HO SUNG + LAO LETTER NO
ໜ	0EDC	0858 0200 0020 0002	LAO LETTER HO NO

ຫມ	0EAB+0EA1	0860 0200 0020 0002	LAO LETTER HO SUNG + LAO LETTER MO
ໝ	0EDD	0860 0200 0020 0002	LAO LETTER HO MO
ຫລ	0EAB+0EA5	0864 0200 0020 0002	LAO LETTER HO SUNG + LAO LETTER LO LOOT
ຫຼ	0EAB+0EBC	0864 0200 0020 0002	LAO LETTER HO SUNG + LAO SEMIVOWEL SIGN LO
ຫວ	0EAB+0EA7	0868 0200 0020 0002	LAO LETTER HO SUNG + LAO LETTER WO
ອ	0EAD	086A 0200 0020 0002	LAO LETTER O
ຮ	0EAE	086C 0200 0020 0002	LAO LETTER HO TAM
← Vowels→			
ະ	0EB0	0000 0202 0020 0002	LAO VOWEL SIGN A
◌ັ+X8/X9	0EB1+X8/X9	0000 0204 0020 0002	LAO VOWEL SIGN MAI KAN + CONSONANTAL
າ	0EB2	0000 0206 0020 0002	LAO VOWEL SIGN AA
◌ິ	0EB4	0000 0208 0020 0002	LAO VOWEL SIGN I
◌ີ	0EB5	0000 020A 0020 0002	LAO VOWEL SIGN II
◌ຶ	0EB6	0000 020C 0020 0002	LAO VOWEL SIGN Y
◌ື	0EB7	0000 0210 0020 0002	LAO VOWEL SIGN YY
◌ຸ	0EB8	0000 0212 0020 0002	LAO VOWEL SIGN U
◌ູ	0EB9	0000 0214 0020 0002	LAO VOWEL SIGN UU
ເXະ	0EC0+X+0EB0	0000 0216 0020 0002	LAO VOWEL SIGN E + MAIN CONSONANT + LAO VOWEL SIGN A
ເ◌ັ+X8/X9	0EC0+0EB1+X8/X9	0000 0218 0020 0002	LAO VOWEL SIGN E + LAO VOWEL SIGN MAIN KAN + CONSONANTAL
ເX	0EC0+X	0000 021A 0020 0002	LAO VOWEL SIGN E + MAIN CONSONANT
ແXະ	0EC1+X+0EB0	0000 021C 0020 0002	LAO VOWEL SIGN EI + MAIN CONSONANT + LAO VOWEL SIGN A
ແ◌ັ+X8/X9	0EC1+0EB1+X8/X9	0000 0220 0020 0002	LAO VOWEL SIGN EI + LAO VOWEL SIGN MAI KAN + CONSONANTAL
ແX	0EC1+X		LAO VOWEL SIGN EI +

		0000 0222 0020 0002	MAIN CONSONANT
ໂXະ	0EC2+X+0EB0	0000 0224 0020 0002	LAO VOWEL SIGN O + MAIN CONSONANT + LAO VOWEL SIGN A
◌ົ	0EBB	0000 0226 0020 0002	LAO VOWEL SIGN MAI KON
ໂX	0EC2+X	0000 0228 0020 0002	LAO VOWEL SIGN O + MAIN CONSONANT
ເXາະ	0EC0+X+0EB2+0EB0	0000 022A 0020 0002	LAO VOWEL SIGN E + MAIN CONSONANT + LAO VOWEL AA + LAO VOWEL SIGN A
◌ໍ	0ECD	0000 022C 0020 0002	LAO NIGGAHITA
Xອ+X8/X9	X+0EAD+X8/X9	0000 022E 0020 0002	MAIN CONSONANT + LAO LETTER O + CONSONANTAL
ເ◌ິ	0EC0+0EB4	0000 0230 0020 0002	LAO VOWEL SIGN E + LAO VOWEL SIGN I
ເ◌ີ	0EC0+0EB5	0000 0232 0020 0002	LAO VOWEL SIGN E + LAO VOWEL SIGN II
ເ◌ັຽ	0EC0+0EB1+0EBD	0000 0234 0020 0002	LAO VOWEL SIGN E + LAO VOWEL SIGN MAI KAN + LAO SEMIVOWEL SIGN NYO
ເXຽ	0EC0+X+0EBD	0000 0236 0020 0002	LAO VOWEL SIGN E + MAIN CONSONANT + LAO SEMI VOWEL SIGN NYO
ຽ+X8/X9	0EBD+X8/X9	0000 0238 0020 0002	LAO SEMI VOWEL SIGN NYO + CONSONANTAL
◌ົວະ	0EBB+0EA7+0EB0	0000 023A 0020 0002	LAO VOWEL SIGN MAI KON + LAO LETTER WO + LAO + VOWEL SIGN A
◌ັວ+X8/X9	0EB1+0EA7+X8/X9	0000 023C 0020 0002	LAO VOWEL SIGN MAI KON + LAO LETTER WO + CONSONANTAL
◌ົວ	0EBB+0EA7	0000 023E 0020 0002	LAO VOWEL SIGN MAI KON + LAO LETTER WO
ເ◌ຶອ	0EC0+0EB6+0EAD	0000 0240 0020 0002	LAO VOWEL SIGN E + LAO VOWEL SIGN Y + LAO LETTER O
ເ◌ືອ	0EC0+0EB7+0EAD	0000 0242 0020 0002	LAO VOWEL SIGN E + LAO VOWEL SIGN YY + LAO LETTER O
Xວ+X8/X9	X+0EA7+X8/X9	0000 0244 0020 0002	MAIN CONSONANT + LAO LETTER WO + CONSONANTAL
ໄX	0EC4+X	0000 0246 0020 0002	LAO VOWEL SIGN AI + MAIN CONSONANT
ໃX	0EC3+X	0000 0248 0020 0002	LAO VOWEL SIGN AY + MAIN CONSONANT
ເ◌ົາ			LAO VOWEL SIGN E +

	0EC0+0EBB+0EB2	0000 024A 0020 0002	LAO VOWEL SIGN MAI KON + LAO VOWEL SIGN AA
◌ຳ	0EB3	0000 024C 0020 0002	LAO VOWEL SIGN AM
◌ໍ+າ	0ECD+0EB2	0000 024C 0020 0002	LAO NIGGAHITA + LAO VOWEL SIGN AA
← Alternate Consonants →			
ກ	0E81	0000 0000 0022 0002	LAO LETTER KO
ງ	0E87	0000 0000 0024 0002	LAO LETTER NGO
ຍ	0E8D	0000 0000 002C 0002	LAO LETTER NYO
ດ	0E94	0000 0000 002E 0002	LAO LETTER DO
ນ	0E99	0000 0000 0030 0002	LAO LETTER NO
ບ	0E9A	0000 0000 0032 0002	LAO LETTER BO
ມ	0EA1	0000 0000 0038 0002	LAO LETTER MO
ວ	0EA7	0000 0000 003C 0002	LAO LETTER WO
← Tone Marks→			
◌່	0EC8	0000 0000 0000 0004	LAO TONE MAI EK
◌້	0EC9	0000 0000 0000 0006	LAO TONE MAI THO
◌໊	0ECA	0000 0000 0000 0008	LAO TONE TI
◌໋	0ECB	0000 0000 0000 0008	LAO TONE MAI CATAWA
← Numerals→			
໐	0ED0	0700 0200 0020 0002	LAO DIGIT ZERO
໑	0ED1	0702 0200 0020 0002	LAO DIGIT ONE
໒	0ED2	0704 0200 0020 0002	LAO DIGIT TWO
໓	0ED3	0706 0200 0020 0002	LAO DIGIT THREE
໔	0ED4	0708 0200 0020 0002	LAO DIGIT FOUR
໕	0ED5	070A 0200 0020 0002	LAO DIGIT FIVE
໖	0ED6	070C 0200 0020 0002	LAO DIGIT SIX
໗	0ED7	070E 0200 0020 0002	LAO DIGIT SEVEN
໘	0ED8	0710 0200 0020 0002	LAO DIGIT EIGHT

໙	0ED9	0712 0200 0020 0002	LAO DIGIT NINE
← Various Symbols→			
◌໌	0ECC	0000 0000 0000 0000	MAI KALAN
ໆ	0EC6	0000 0000 0000 0000	MAI SUM
ຯ	0EAF	0000 0000 0000 0000	MAI SUM

5.2.2.2. Script Based Sorting

The collation elements for language based sorting are given in Table 5.9.

Table 5.9. Lao Collation Elements for Script Based Sorting

Glyph	Unicode	Collation Elements	Unicode Name
← Consonants→			
ກ	0E81	0820 0200 0020 0002	LAO LETTER KO
ຂ	0E82	0822 0200 0020 0002	LAO LETTER KHO SUNG
ຄ	0E84	0824 0200 0020 0002	LAO LETTER KHO TAM
ງ	0E87	0826 0200 0020 0002	LAO LETTER NGO
ຈ	0E88	0828 0200 0020 0002	LAO LETTER CO
ສ	0EAA	082A 0200 0020 0002	LAO LETTER SO SUNG
ຊ	0E8A	082C 0200 0020 0002	LAO LETTER SO TAM
ຍ	0E8D	082E 0200 0020 0002	LAO LETTER NYO
ດ	0E94	0830 0200 0020 0002	LAO LETTER DO
ຕ	0E95	0832 0200 0020 0002	LAO LETTER TO
ຖ	0E96	0834 0200 0020 0002	LAO LETTER THO SUNG
ທ	0E97	0836 0200 0020 0002	LAO LETTER THO TAM
ນ	0E99	0838 0200 0020 0002	LAO LETTER NO
ບ	0E9A	083A 0200 0020 0002	LAO LETTER BO
ປ	0E9B	083C 0200 0020 0002	LAO LETTER PO
ຜ			

	0E9C	083E 0200 0020 0002	LAO LETTER PHO SUNG
ຝ	0E9D	0840 0200 0020 0002	LAO LETTER FO TAM
ພ	0E9E	0842 0200 0020 0002	LAO LETTER PHO TAM
ຟ	0E9F	0844 0200 0020 0002	LAO LETTER FO SUNG
ມ	0EA1	0846 0200 0020 0002	LAO LETTER MO
ຢ	0EA2	0848 0200 0020 0002	LAO LETTER YO
ຣ	0EA3	084A 0200 0020 0002	LAO LETTER LO LING
ລ	0EA5	084E 0200 0020 0002	LAO LETTER LO LOOT
ວ	0EA7	0850 0200 0020 0002	LAO LETTER WO
ຫ	0EAB	0852 0200 0020 0002	LAO LETTER HO SUNG
ຫຼ	0EAB+0EBC	0866 0200 0020 0002	LAO LETTER HO SUNG + LAO SEMIVOWEL SIGN LO
ອ	0EAD	086A 0200 0020 0002	LAO LETTER O
ຮ	0EAE	086C 0200 0020 0002	LAO LETTER HO TAM
ໜ	0EDC	0870 0200 0020 0002	LAO LETTER HO NO
ໝ	0EDD	0872 0200 0020 0002	LAO LETTER HO MO
← Vowels→			
ະ	0EB0	0000 0202 0020 0002	LAO VOWEL SIGN A
າ	0EB2	0000 0206 0020 0002	LAO VOWEL SIGN AA
◌ິ	0EB4	0000 0208 0020 0002	LAO VOWEL SIGN I
◌ີ	0EB5	0000 020A 0020 0002	LAO VOWEL SIGN II
◌ຶ	0EB6	0000 020C 0020 0002	LAO VOWEL SIGN Y
◌ື	0EB7	0000 0210 0020 0002	LAO VOWEL SIGN YY
◌ຸ	0EB8	0000 0212 0020 0002	LAO VOWEL SIGN U
◌ູ	0EB9	0000 0214 0020 0002	LAO VOWEL SIGN UU
ເ	0EC0	0000 0216 0020 0002	LAO VOWEL SIGN
ແ	0EC1	0000 0222 0020 0002	LAO VOWEL SIGN EI
ໂ	0EC2	0000 0224 0020 0002	LAO VOWEL SIGN O
◌ໍ	0ECD	0000 022C 0020 0002	LAO NIGGAHITA

ໄ	0EC4	0000 0246 0020 0002	LAO VOWEL SIGN AI
ໃ	0EC3	0000 0248 0020 0002	LAO VOWEL SIGN AY
◌ັ	0EB1	0000 024A 0020 0002	LAO VOWEL SIGN MAI KAN
◌ົ	0EBB	0000 024C 0020 0002	LAO VOWEL SIGN MAI KON
◌ຽ	0EBD	0000 0250 0020 0002	LAO SEMI VOWEL SIGN NYO
ວ	0EA7	0000 0252 0020 0002	LAO LETTER WO
ອ	0EAD	0000 0254 0020 0002	LAO LETTER O
← Consonantal→			
ກ	0E81	0000 0000 0022 0002	LAO LETTER KO
ງ	0E87	0000 0000 0024 0002	LAO LETTER NGO
ຈ	0E88	0000 0000 0026 0002	LAO LETTER CO
ສ	0EAA	0000 0000 0028 0002	LAO LETTER SO SUNG
ຊ	0E8A	0000 0000 002A 0002	LAO LETTER SO TAM
ຍ	0E8D	0000 0000 002C 0002	LAO LETTER NYO
ດ	0E94	0000 0000 002E 0002	LAO LETTER DO
ນ	0E99	0000 0000 0030 0002	LAO LETTER NO
ບ	0E9A	0000 0000 0032 0002	LAO LETTER BO
ພ	0E9E	0000 0000 0034 0002	LAO LETTER PHO TAM
ຟ	0E9F	0000 0000 0036 0002	LAO LETTER FO SUNG
ມ	0EA1	0000 0000 0038 0002	LAO LETTER MO
ລ	0EA5	0000 0000 003A 0002	LAO LETTER LO LOOT
ວ	0EA7	0000 0000 003C 0002	LAO LETTER WO
← Tone Marks→			
◌່	0EC8	0000 0000 0000 0004	LAO TONE MAI EK
◌້	0EC9	0000 0000 0000 0006	LAO TONE MAI THO
◌໊	0ECA	0000 0000 0000 0008	LAO TONE TI
◌໋	0ECB	0000 0000 0000 0008	LAO TONE MAI CATAWA

← **Numerals**→			
໐	0ED0	0700 0200 0020 0002	LAO DIGIT ZERO
໑	0ED1	0702 0200 0020 0002	LAO DIGIT ONE
໒	0ED2	0704 0200 0020 0002	LAO DIGIT TWO
໓	0ED3	0706 0200 0020 0002	LAO DIGIT THREE
໔	0ED4	0708 0200 0020 0002	LAO DIGIT FOUR
໕	0ED5	070A 0200 0020 0002	LAO DIGIT FIVE
໖	0ED6	070C 0200 0020 0002	LAO DIGIT SIX
໗	0ED7	070E 0200 0020 0002	LAO DIGIT SEVEN
໘	0ED8	0710 0200 0020 0002	LAO DIGIT EIGHT
໙	0ED9	0712 0200 0020 0002	LAO DIGIT NINE
← **Various Symbols**→			
◌໌	0ECC	0000 0000 0000 0000	MAI KALAN
ໆ	0EC6	0000 0000 0000 0000	MAI SUM
ຯ	0EAF	0000 0000 0000 0000	MAI SUM

Results

Data sorted by different strategies gives different output sequences. Sample output sequences for each technique are given in Tables 5.10 and 5.11.

Table 5.10. Input and Corresponding Sorted Output for Lao Using Language Based Strategy

Sample Input		**Sample Output**	
ເງິນແຮຖົງ	ກົກເສົາ	ກະ ໂຄງການ	ກົ້ນຫນັກ
ກາ	ກົ້ນຂີ້	ກະຕື້ລືລົ້ນ	ກໍຈິງຢູ່ແລ້ວ
ຈົດທະບຽນ	ກົ້ນຂີ້ທັ່ງ	ກະຕື້ລືລົ້ນ	ກອກນ້ຳ
ສຽງມ້າແຫມ	ຈົດທະບຽນການ	ກະແຕະ	ກອງກົ້ນ
ສຽງທອງ	ຄ້າ	ກະແຕະ	ກອງ ໂຈນ
ເກືອບຫມົດ	ງານຂຶ້ນເຮືອນໃຫ	ກະແຕ	ກອງສອດແນມ

ສີກຸຫລາບ	ມ່	ກະເຕາະກະແຕ	ເກີດຈາກ
ເງິນຮາງ	ກາກະບາດ	ະ	ເກີດໄພ
ກ້ວຍ	ສຽງຟ້າຮ້ອງ	ກັບຄືນມາ	ເກີດມາ
ໄກເທົ່າໃດ	ແກ້ວໂກເມນ	ກາ	ກົວ
ກະແຕ	ກົ້ນຂວດ	ກາກະບາດ	ກົ່ວ
ເກີດມາ	ກົ້ນຫນັກ	ກາຄຳຊອບ	ເກືອບຫມົດ
ກາຄຳຊອບ	ກໍຈິງຢູ່ແລ້ວ	ກາໂຕລິກ	ກວຍ
ເກົ່າ	ເຄື້ອງວັດຄວາມໄ	ກ້າ	ກ້ວຍ
ກະໂຄງການ	ວ	ກ້າ	ກ້ວຍມືນາງ
ກ້າ	ກອກນ້ຳ	ກ້າ	ໄກເທົ່າໃດ
ກ້າ	ກົກຂາ	ກ້າກັ່ນ	ໄກປານໃດ
ກິນິນ	ກ້າແກ່ນ	ກ້າແກ່ນ	ໃກ້ຊິດກັນ
ກາໂຕລິກ	ກັບຄືນມາ	ກິນິນ	ເກົາ
ກິຣິຍາ	ເກີດຈາກ	ກິຣິຍາ	ເກົ່າ
ກື້ເຕາ	ເກີດໄພ	ກື້ເຕາ	ເຄື້ອງວັດຄວາມໄ
ກື້ງຕາໃສ່	ກອງສອດແນມ	ກື້ງຕາໃສ່	ວ
ກ້າກັ່ນ	ຈັກກະພັດນິຍົມ	ກຸຫລາບ	ເຄື້ອງວັດຄວາມ
ກ້າ	ກະເຕາະກະແຕະ	ກູລີ	ຮ້ອນເຢັນ
ກູລີ	ກອງກົ້ນ	ເກັດປາ	ງານຂຶ້ນເຮືອນໃຫ
ກຸຫລາບ	ກົ່ວ	ເກັບ	ມ່
ເກັດປາ	ກວຍ	ເກັບກ່ຽວ	ເງິນຮາງ
ສຽງແຫບ	ກ້ວຍມືນາງ	ເກັບກ່ຽວ	ເງິນແຮຖ່ງ
ເກັບ	ໃກ້ຊິດກັນ	ເກັບພາສີ	ຈັກກະພັດນິຍົມ
ກະຕື້ລືລົ້ນ	ກະແຕະ	ແກ້ວໂກເມນ	ຈົດທະບຽນ
ເກັບກ່ຽວ	ໄກປານໃດ	ແກ້ວນາເຝານເ	ຈົດທະບຽນການ
ແກ້ວນາເຝານເ	ເກົາ	ນື້ອແຂງ	ຄ້າ
ນື້ອແຂງ	ກອງໂຈນ	ກົກ	ສີກຸຫລາບ
ກົກ	ຊັກອອກພູດນິ່ງ	ກົກຂາ	ສີແກ່
ເກັບກ່ຽວ	ກົວ	ກົກແຂນ	ສຽງທອງ
ເກັບພາສີ	ສີແກ່	ກົກເສົາ	ສຽງປົກກະຕິ
ກະຕື້ລືລົ້ນ	ເຄື້ອງວັດຄວາມ	ກົ້ນຂີ້	ສຽງຟ້າຮ້ອງ
ກົກແຂນ	ຮ້ອນເຢັນ	ກົ້ນຂີ້ທັ່ງ	ສຽງມ້າແຫມ

	ກະແຕະ ສຽງປີກກະຕິ	ກົ້ນຂວດ	ສຽງແຫບ ຊັກອອກພູດນີ້

Table 5.11. Input and Corresponding Sorted Output for Lao Using Script Based Strategy

Sample Input		Sample Output	
ເງິນແຮຖົງ	ກີກເສົາ	ກະໂຄງການ	ໄກປານໃດ
ກາ	ກົ້ນຂີ້	ກະຕີ້ລີລົ້ນ	ໃກ້ຊິດກັນ
ຈິດທະບຽນ	ກົ້ນຂີ້ທັ່ງ	ກະຕີ້ລີລົ້ນ	ກັບຄືນມາ
ສຽງມ້າແຫມ	ຈິດທະບຽນການ	ກະເຕາະກະແຕ	ກີກ
ສຽງທອງ	ຄ້າ	ະ	ກີກຂາ
ເກືອບຫມົດ	ງານຂຶ້ນເຮືອນໃຫ	ກະແຕ	ກີກແຂນ
ສີກຸຫລາບ	ມ່	ກະແຕະ	ກີກເສົາ
ເງິນຮາງ	ກາກະບາດ	ກະແຕະ	ກົ້ນຂີ້
ກ້ວຍ	ສຽງຟ້າຮ້ອງ	ກາ	ກົ້ນຂີ້ທັ່ງ
ໄກເທົ່າໃດ	ແກ້ວໂກເມນ	ກາກະບາດ	ກົ້ນຂວດ
ກະແຕ	ກົ້ນຂວດ	ກາຄຳຊອບ	ກົ້ນຫນັກ
ເກີດມາ	ກົ້ນຫນັກ	ກາໂຕລິກ	ກົວ
ກາຄຳຊອບ	ກໍຈິງຢູ່ແລ້ວ	ກ້າ	ກົ່ວ
ເກົ່າ	ເຄື້ອງວັດຄວາມໄ	ກ້າ	ກວຍ
ກະໂຄງການ	ວ	ກ້າ	ກ້ວຍ
ກ້າ	ກອກນ້ຳ	ກ້າແກ່ນ	ກ້ວຍມືນາງ
ກ້າ	ກີກຂາ	ກ້າກັ່ນ	ກອກນ້ຳ
ກິນິນ	ກ້າແກ່ນ	ກິນິນ	ກອງກົ້ນ
ກາໂຕລິກ	ກັບຄືນມາ	ກິຣິຍາ	ກອງໂຈນ
ກິຣິຍາ	ເກີດຈາກ	ກີ້ເຕາ	ກອງສອດແນມ
ກີ້ເຕາ	ເກີດໄພ	ກີ້ງຕາໃສ່	ເຄື້ອງວັດຄວາມໄ
ກີ້ງຕາໃສ່	ກອງສອດແນມ	ກຸຫລາບ	ວ
ກ້າກັ່ນ	ຈັກກະພັດນິຍົມ	ກູລິ	ເຄື້ອງວັດຄວາມ
ກ້າ	ກະເຕາະກະແຕະ	ເກີດຈາກ	ຮ້ອນເຢັນ
ກູລິ	ກອງກົ້ນ	ເກີດໄພ	ງານຂຶ້ນເຮືອນໃຫ

ກຸຫລາບ	ກົ່ວ	ເກີດມາ	ມ່
ເກັດປາ	ກວຍ	ເກືອບຫມົດ	ເງິນຮາງ
ສຽງແຫບ	ກ້ວຍມືນາງ	ເກັດປາ	ເງິນແຮຖົງ
ເກັບ	ໃກ້ຊິດກັນ	ເກັບ	ຈັກກະພັດນິຍົມ
ກະຕື້ລືລົ້ນ	ກະແຕະ	ເກັບກ່ຽວ	ຈົດທະບຽນ
ເກັບກ່ຽວ	ໄກປານໃດ	ເກັບກ່ຽວ	ຈົດທະບຽນການ
ແກ້ວນາເຜານເ	ເກົາ	ເກັບພາສີ	ຄ້າ
ນື້ອແຂງ	ກອງໂຈນ	ເກົາ	ສີກຸຫລາບ
ກົກ	ຊັກອອກພູດນຶ່ງ	ເກົ່າ	ສີແກ່
ເກັບກ່ຽວ	ກົວ	ແກ້ວໂກເມນ	ສຽງທອງ
ເກັບພາສີ	ສີແກ່	ແກ້ວນາເຜານເ	ສຽງປົກກະຕິ
ກະຕື້ລືລົ້ນ	ເຄື້ອງວັດຄວາມ	ນື້ອແຂງ	ສຽງຟ້າຮ້ອງ
ກົກແຂນ	ຮ້ອນເຢັນ	ກໍຈິງຢູ່ແລ້ວ	ສຽງມ້າແຫມ
	ກະແຕະ	ໄກເທົ່າໃດ	ສຽງແຫບ
	ສຽງປົກກະຕິ		ຊັກອອກພູດນຶ່

Conclusion

Lao presents one of the most challenging scenarios for collation. First, Lao text does not have spaces so processing is required to segment text into words (not discussed in detail in this chapter; much work has been done on this for Thai, e.g. see [23, 24, 25]). Once the word sequence is available, words are required to be syllabified and individual characters need to be tagged for different roles depending on the context (details of this process are discussed in [20]). Then re-ordering and normalization need to be done. Finally, depending on collation strategy, which could be based on language or script, collation elements need to be assigned. Within the syllable, Lao sorts at four levels, with nuclear consonants getting the primary weight, vowels getting the secondary weight, non-nuclear consonants getting the tertiary weight and tone marks getting the quaternary level weight. The sort keys generated are also at syllable level (and not at word level). Thus, the Unicode collation algorithm [2] discussed in the second chapter needs to be modified to do a sequence of comparisons of sort keys generated by syllables from words.

Though the current work has been tested, much more work needs to be done in this area. Standards also need to be defined by relevant organizations.

6. Mongolian

Mongolian is an Altaic language spoken in Mongolia, China and Russian Federation. Today about 8 million people in the world speak Mongolian. Most of that are approximately 2.7 million in Mongolia and 3.38 million in Inner Mongolia in China [37, 38]. Khalkha or Halha dialects of Mongolian is the national language of Mongolia [37].

6.1. Writing System

Mongolian has shown a varied history of writing. Early Mongolian was written in a script adapted from Old Sogdo-Uighur script in early thirteenth century. As Mongolian derived from Uighur script which originated from Aramaic script (of Semitic origin), it was initially written in a right-to-left direction. However, later the system was rotated by 90 degrees counter clockwise and currently the script is written in top down direction from left-top-right, a unique feature of this script [39]. However, over next two centuries Chinese, Arabic and Tibetan scripts were also used to write the language. In 1930's Cyrillic script was increasingly used, and on 1st, Jan, 1946 it was formally adopted by the Mongolian government. It is still being used to write Mongolian language in Mongolia. There have been attempts to restore Traditional Mongolian script, e.g. by the government in 1994 [39]. Though currently both scripts are used in Mongolia, Cyrillic use is more widespread. Traditional Mongolian script is mostly being used in Inner Mongolia in China to write Mongolian. The Cyrillic and Traditional Mongolian scripts do not have clear correspondence. The current work is focused on the collation of Mongolian language using Cyrillic script.

6.1.1. Character Set

Cyrillic script has been derived from Greek script and has been traditionally used to write Slavic languages, including Russian. The Mongolian character set is slightly modified Cyrillic alphabet by adding two vowels such as (ө, ү). Each character has a capital and a small letter (shown in figure below) and it uses the numerals 0, 1, 2, 3…, 9.

А Б В Г Д Е Ё Ж З И Й К Л М Н О Ө П
Р С Т У Ү Ф Х Ц Ч Ш Щ Ъ Ы Ь Э Ю Я

Capital Letters

а б в г д е ё ж з и й к л м н о ө п
р с т у ү ф х ц ч ш щ ъ ы ь э ю я

Small Letters

Figure 6.1. Mongolian Character Set in Cyrillic Script

6.1.2. Script Details

Cyrillic is written from left to right. Words are separated by spaces and letters are cased as capital and small letters.

6.1.2.1. Case

In Mongolian, all the characters have upper and lower case variants, with the exception of Palochka [4]. For example Cyrillic letter Zhe has the upper case form Ж (U+0416) and the lower case form ж (U+0436). The characters with upper case are sorted before the ones with lower case.

6.2. Collation

Mongolian uses the conventional ordering of Cyrillic script and the three levels of collation associated with it. Numerals and letters are sorted at primary level, diacritics are sorted at secondary level, and case is handled at the tertiary level.

6.2.1. Text Processing

6.2.1.1. Normalization

Mongolian has few characters which can be encoded in multiple ways using Unicode. This is possible due to separate encoding of some marks in addition to encoding of composite forms. Some examples are shown in Table 6.1. below.

Table 6.1. Examples of Normalization in Mongolian using Cyrillic Script

Decomposed Form	Unicodes of Decomposed Form	Equivalent Composed Form	Unicode of Composed Form
Ё	0415 0308	Ё	0401
ё	0435 0308	ё	0451
Й	0418 0306	Й	0419

6.2.2. Unicode Collation Elements

Following collation elements give correct ordering of Mongolian script. The results are based on the order of words given in [40].

Table 6.2. Collation Elements for Mongolian Language Using Cyrillic Script

Glyph	Unicode	Collation Elements	Unicode Name
← Numerals →			
0	0030	00A0 0020 0002	DIGIT ZERO
1	0031	00A1 0020 0002	DIGIT ONE
2	0032	00A2 0020 0002	DIGIT TWO
3	0033	00A3 0020 0002	DIGIT THREE
4	0034	00A4 0020 0002	DIGIT FOUR
5	0035	00A5 0020 0002	DIGIT FIVE
6	0036	00A6 0020 0002	DIGIT SIX
7	0037	00A7 0020 0002	DIGIT SEVEN
8	0038	00A8 0020 0002	DIGIT EIGHT
9	0039	00A9 0020 0002	DIGIT NINE
←Consonants and Vowels→			
А	0410	0E29 0020 0002	CYRILLIC CAPITAL LETTER A
а	0430	0E29 0020 0008	CYRILLIC SMALL LETTER A
Б	0411	0E2A 0020 0002	CYRILLIC CAPITAL LETTER BE
б	0431	0E2A 0020 0008	CYRILLIC SMALL LETTER BE
В	0412	0E2B 0020 0002	CYRILLIC CAPITAL LETTER VE
в	0432	0E2B 0020 0008	CYRILLIC SMALL LETTER VE
Г	0413	0E2C 0020 0002	CYRILLIC CAPITAL LETTER GHE
г	0433	0E2C 0020 0008	CYRILLIC SMALL LETTER GHE
Д	0414	0E2D 0020 0002	CYRILLIC CAPITAL LETTER DE
д	0434	0E2D 0020 0008	CYRILLIC SMALL LETTER DE
Е	0415	0E2E 0020 0002	CYRILLIC CAPITAL LETTER IE
е	0435	0E2E 0020 0008	CYRILLIC SMALL LETTER IE
Ё	0401	0E2F 0020 0002	CYRILLIC CAPITAL LETTER IO
Е ̈	0415 0308	0E2F 0020 0002	CYRILLIC CAPITAL LETTER IO
ё	0451	0E2F 0020 0008	CYRILLIC SMALL LETTER IO
ё	0435 0308	0E2F 0020 0008	CYRILLIC SMALL LETTER IO
Ж	0416	0E30 0020 0002	CYRILLIC CAPITAL LETTER ZHE
ж	0436	0E30 0020 0008	CYRILLIC SMALL LETTER ZHE
З	0417	0E31 0020 0002	CYRILLIC CAPITAL LETTER ZE
з	0437	0E31 0020 0008	CYRILLIC SMALL LETTER ZE
И	0418	0E32 0020 0002	CYRILLIC CAPITAL LETTER I
и	0438	0E32 0020 0008	CYRILLIC SMALL LETTER I
Й	0419	0E33 0020 0002	CYRILLIC CAPITAL LETTER SHORT I
Й	0418 0306	0E33 0020 0002	CYRILLIC CAPITAL LETTER SHORT I
й	0439	0E33 0020 0008	CYRILLIC SMALL LETTER SHORT I
й	0438 0306	0E33 0020 0008	CYRILLIC SMALL LETTER SHORT I
К	041A	0E34 0020 0002	CYRILLIC CAPITAL LETTER KA
к	043A	0E34 0020 0008	CYRILLIC SMALL LETTER KA
Л	041B	0E35 0020 0002	CYRILLIC CAPITAL LETTER EL
л	043B	0E35 0020 0008	CYRILLIC SMALL LETTER EL
М	041C	0E36 0020 0002	CYRILLIC CAPITAL LETTER EM
м	043C	0E36 0020 0008	CYRILLIC SMALL LETTER EM

Н	041D	0E37 0020 0002	CYRILLIC CAPITAL LETTER EN
н	043D	0E37 0020 0008	CYRILLIC SMALL LETTER EN
О	041E	0E38 0020 0002	CYRILLIC CAPITAL LETTER O
о	043E	0E38 0020 0008	CYRILLIC SMALL LETTER O
Ө	04E8	0E39 0020 0002	CYRILLIC CAPITAL LETTER BARRED O
ө	04E9	0E39 0020 0008	CYRILLIC SMALL LETTER BARRED O
П	041F	0E3A 0020 0002	CYRILLIC CAPITAL LETTER PE
п	043F	0E3A 0020 0008	CYRILLIC SMALL LETTER PE
Р	0420	0E3B 0020 0002	CYRILLIC CAPITAL LETTER ER
р	0440	0E3B 0020 0008	CYRILLIC SMALL LETTER ER
С	0421	0E3C 0020 0002	CYRILLIC CAPITAL LETTER ES
с	0441	0E3C 0020 0008	CYRILLIC SMALL LETTER ES
Т	0422	0E3E 0020 0002	CYRILLIC CAPITAL LETTER TE
т	0442	0E3E 0020 0008	CYRILLIC SMALL LETTER TE
У	0423	1350 0020 0002	CYRILLIC CAPITAL LETTER U
у	0443	1350 0020 0008	CYRILLIC SMALL LETTER U
Ү	04AE	1353 0020 0002	CYRILLIC CAPITAL LETTER STRAIGHT U
ү	04AF	1353 0020 0008	CYRILLIC SMALL LETTER STRAIGHT U
Ф	0424	1356 0020 0002	CYRILLIC CAPITAL LETTER EF
ф	0444	1356 0020 0008	CYRILLIC SMALL LETTER EF
Х	0425	1359 0020 0002	CYRILLIC CAPITAL LETTER HA
х	0445	1359 0020 0008	CYRILLIC SMALL LETTER HA
Ц	0426	135C 0020 0002	CYRILLIC CAPITAL LETTER TSE
ц	0446	135C 0020 0008	CYRILLIC SMALL LETTER TSE
Ч	0427	135F 0020 0002	CYRILLIC CAPITAL LETTER CHE
ч	0447	135F 0020 0008	CYRILLIC SMALL LETTER CHE
Ш	0428	1360 0020 0002	CYRILLIC CAPITAL LETTER SHA
ш	0448	1360 0020 0008	CYRILLIC SMALL LETTER SHA
Щ	0429	1363 0020 0002	CYRILLIC CAPITAL LETTER SHCHA
щ	0449	1363 0020 0008	CYRILLIC SMALL LETTER SHCHA
Ъ	042A	1366 0020 0002	CYRILLIC CAPITAL LETTER HARD SIGN
ъ	044A	1366 0020 0008	CYRILLIC SMALL LETTER HARD SIGN
Ы	042B	1369 0020 0002	CYRILLIC CAPITAL LETTER YERU
ы	044B	1369 0020 0008	CYRILLIC SMALL LETTER YERU
Ь	042C	136C 0020 0002	CYRILLIC CAPITAL LETTER SOFT SIGN
ь	044C	136C 0020 0008	CYRILLIC SMALL LETTER SOFT SIGN
Э	042D	136F 0020 0002	CYRILLIC CAPITAL LETTER E
э	044D	136F 0020 0008	CYRILLIC SMALL LETTER E
Ю	042E	1370 0020 0002	CYRILLIC CAPITAL LETTER YU
ю	044E	1370 0020 0008	CYRILLIC SMALL LETTER YU
Я	042F	1373 0020 0002	CYRILLIC CAPITAL LETTER YA
я	044F	1373 0020 0008	CYRILLIC SMALL LETTER YA

6.2.3. Results

Table 6.3. shows output obtained by sorting a sample input using the collation elements given in Table 6.2.

Table 6.3. Input and Corresponding Sorted Output for Mongolian

Input		Output	
Яион	Маяг	Аагим	ёслогч
ганц	Каир	аагим	ёст
бүч	Ёстой	Аагтай	Ёстой
Бзл	ааЖуу	Аагтай	Ёстой
бзл	ИГ	аагтай	ёстой
Аагим	аагтай	аагтай	ёстой
ганха	ИД	ааЖуу	ИГ
Тойн	егее	аажуу	ИД
год	Аагтай	Бзл	Кабин
ёстой	Цунх	бзл	Каир
дзздзх	Яри	бүч	Маяг
аагтай	Метр	ганха	Метр
дззр	Тожгор	ганц	Тожгор
ёстой	аагим	год	Тойн
Кабин	аажуу	дзздзх	Цунх
Аагтай	еГЕЕ	дззр	Цуца
ёслогч	Ёстой	еГЕЕ	Яион
ёслогч	Цуца	егее	Яри
ёст		ёслогч	

6.3. *Conclusion*

Mongolian is a simple case of collation. It is very similar to that of other Latin and Cyrillic based languages. Letters are sorted at primary level, marks are sorted at secondary level and case is sorted at tertiary level. There are no exceptions to this process. Some pre-processing is required before collation can be done to normalize multiple encodings.

7. Sindhi

Sindhi is an Indo-Aryan language spoken by 18.5 million people in Pakistan and 2.8 million people in India. It is a state language in both countries [41]. Sindhi is written using extended Arabic script in Naskh style in Pakistan and in Devanagari Script in India. Current work is based on the Arabic script based system.

7.1. *Writing System*

7.1.1. Character Set

Sindhi character set, based on Perso-Arabic writing system, was introduced around 1852 [42]. It is written from right-to-left and introduces additional characters to cater to additional features of Sindhi language. Sindhi character set has 52 letters representing the consonants and long vowels. These are listed in Figure 7.1.

ا ب ٻ ڀ ت ٿ ٽ ٺ ث پ ڦ ج ڄ جھ ڃ چ ڇ ح خ د ڌ ڏ ڊ ڍ
ذ ر ڙ ز س ش ص ض ط ظ ع غ ف ق ڪ ک گ ڳ گھ ڱ ل م ن
ڻ و ھ ء ی

Figure 7.1. Sindhi Character Set

Short vowels and some additional vocalic and consonantal features are also represented through diacritical marks in Sindhi [43]. These are listed in Figure 7.2. The diacritics (also known as *aerab*) are optionally used in writing. Native speakers use their inherent knowledge of the language to determine the pronunciation when the diacritical vowel marking are missing.

بَ بِ بُ بْ بً بٰ بّ

Figure 7.2. Sindhi Diacritics

Sindhi also has honorific marks which are used to show respect, and are used with proper names. These honorifics are shown in Figure 7.3.

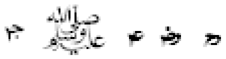

Figure 7.3. Honorific Marks in Sindhi

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

٩ ٨ ٧ ۶ ۵ ۴ ٣ ٢ ١ ٠

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 ۾ ١٢٣ 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 ى 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
ﻻ	FEFB	ل ا	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+0647[1]) 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.

ج + ھ =جھ
گ + ھ = گھ

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 →			
۰	06F0	0E29 0020 0002	ARABIC-INDIC DIGIT ZERO
۱	06F1	0E2A 0020 0002	ARABIC-INDIC DIGIT ONE
۲	06F2	0E2B 0020 0002	ARABIC-INDIC DIGIT TWO
۳	06F3	0E2C 0020 0002	ARABIC-INDIC DIGIT THREE
۴	06F4	0E2D 0020 0002	ARABIC-INDIC DIGIT FOUR
۵	06F5	0E2E 0020 0002	ARABIC-INDIC DIGIT FIVE
۶	06F6	0E2F 0020 0002	ARABIC-INDIC DIGIT SIX
۷	06F7	0E30 0020 0002	ARABIC-INDIC DIGIT SEVEN
۸	06F8	0E31 0020 0002	ARABIC-INDIC DIGIT EIGHT
۹	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
ٻ	067B	1356 0020 0002	ARABIC LETTER BEEH
ڀ	0680	1359 0020 0002	ARABIC LETTER BEHEH

[1] Not recommended for use for Sindhi.

ت	062A	135C 0020 0002	ARABIC LETTER TEH
ٿ	067F	135F 0020 0002	ARABIC LETTER TEHEH
ٽ	067D	1360 0020 0002	ARABIC LETTER THE WITH THREE DOTS ABOVE DOWNWARDS
ٺ	067A	1363 0020 0002	ARABIC LETTER TTEHEH
ث	062B	1366 0020 0002	ARABIC LETTER THEH
پ	067E	1369 0020 0002	ARABIC LETTER PEH
ڦ	06A6	136C 0020 0002	ARABIC LETTER PEHEH
ج	062C	136F 0020 0002	ARABIC LETTER JEEM
ڄ	0684	1370 0020 0002	ARABIC LETTER DYEH
جھ	062C 06BE	1373 0020 0002	ARABIC LETTER JEEM + ARABIC LETTER HEH DOCHASHMEE
ڃ	0683	1376 0020 0002	ARABIC LETTER NYEH
چ	0686	1379 0020 0002	ARABIC LETTER TCHEH
ڇ	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
ڌ	068C	1386 0020 0002	ARABIC LETTER DAHAL
ڏ	068F	1389 0020 0002	ARABIC LETTER DAL WITH THREE DOTS ABOVE DOWNWARD
ڊ	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
ڙ	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
ض	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
ڪ	06AA	13B6 0020 0002	ARABIC LETTER SWASH KAF
ک	06A9	13B9 0020 0002	ARABIC LETTER KEHEH
گ	06AF	13BC 0020 0002	ARABIC LETTER GAF

ڳ	06B3	13BF 0020 0002	ARABIC LETTER GUEH
گھ	06AF 06BE	13C0 0020 0002	ARABIC LETTER GAF + ARABIC LETTER HEH DOCHASHMEE
ڱ	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
ڻ	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
◌ٰ	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	
ڊِرَ	آريا ڙڻ	ڦ و	اراوِڻ
رانِيئِڻُ	آريڪڻ	ڦ ي	شَهوت
سقَر	آڙچڻ	ڦ ايڻ	صبرٍ
سَقِرو	ڦَ	ڄُٹايڻ	ضِيقٍ
سُقِرو	ڦِ	ڇتو	طوفانُ
شَهوت	ڦٍ	ڇِتو	عظمى
صبرٍ	ڦ ائيڻ	ڇٍِ ائيڻ	قِشِمِشِ
ضِيقٍ	ڦ و	ڇِ ڻ	امولَ
طوفانُ	ڦ ى	حادِثو	ڪاتو
عظمى	ڄٹائيڻ	حادُيِثو	ڪابو
قِشِمِش	ڇٍتو	ڏوپِ	آريڪڻ
اراوِڻ	ڇتو	ڏوپٍ	گنا
امولَ	ڇٍِ ائيڻ	ڏوپُ	گهَلاٹو
ڪابو	ڇِ ڻ	ڏاه	گهر
ڪاتو	حادِثو	ڏاهِ	لا
گنا	حادِيِثو	ڏاهِى	لَگِنُ
گهِر	ڏوپِ	ڊِڍٍ	مَنَارٍڻ
گهَلاٹو	ڏوپٍ	ڊِڍِ	وِاتٍ
لا	ڏوپ	ڊِڍِرُ	هٽُ
لَگِنُ	ڏاهٍ	ڊِر	يتيم
مَنَارٍڻ	ڏاهٌ	رانِيئِڻُ	آريا ڙڻ
وِاتٍ	ڏاهِى	سِقَر	آڙچڻ
هٽُ	ڊِڍٍ	سَقِرو	ڦِ
يتيم	ڊِڍِ	سُقرو	ڦَ

	ڍَڍَرُ		ۋُ

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.

8. Sinhala

Sinhala is an Indo-Aryan language, spoken in Sri Lanka by about 13 million people, and also known as Sinhalese and Singhalese [45]. Sinhala is the one of the national languages of Sri Lanka. It has a significantly different written and spoken form, with literary form influenced by Sanskrit [46, 18].

Sinhala script is a descendant of Brahmi script and was formed between 7th -8th century [18] and has a syllabic writing system, like other Indic scripts. However, the system is unique because it has distinct rounded forms with no top-line, similar to South Indian scripts, latter used to write Dravidian languages.

8.1. *Writing System*

8.1.1. Character Set

Though various sources list slightly different number of consonants and vowels (e.g. [5, 47, 18], 49, 50, 51]), a latest work on Sinhala collation [48] fixes the count at 41 consonants, 18 vowels, 2 semi-consonants and symbols for dependent vowels. The dependent vowels are known as vowel-strokes or *Pili* in Sinhala. Multiple letters may be used to represent some of these sounds [48] and thus the total number of letters include 35 symbols for consonants, 6 symbols for nasal consonants, and 18 symbols for independent vowels and 17 symbols for dependent vowels (as a dependent vowel is inherent and does not require an explicit symbol) [5, 51]. These are shown in Figures 8.1 and 8.2.

අ ආ ඇ ඈ ඉ ඊ උ ඌ ඍ ඎ ඏ ඐ එ ඒ ඓ ඔ ඕ ඖ

Independent Vowels

◌ා ◌ැ ◌ෑ ◌ි ◌ී ◌ු ◌ූ ◌ෘ ◌ෲ ◌ෟ ◌ෳ ෙ◌ ේ◌ ෛ◌ ො◌ ෝ◌ ෞ◌

Dependent Vowels

Figure 8.1. Sinhala Vowels [47]

ක ඛ ග ඝ ඞ ඟ ච ඡ ජ ඣ ඤ ඥ ඦ ට ඨ ඩ ඪ ණ ඬ ත
ථ ද ධ න ඳ ප ඵ බ භ ම ඹ ය ර ල ව ශ ෂ ස හ ළ ෆ

Figure 8.2. Sinhala Consonants [47]

Other than consonants and vowels there are some special characters or modifiers including Virama (also called Al-Lakuna), Visarga and Anusvara, given in Table 8.1.

Table 8.1. Special Marks in Sinhala

Name	Glyph	Usage
Virama / Al-Lakuna	◌්	ක්
Anusvara	ං	කං
Visarga	ඃ	කඃ

Al-Lakuna is discussed in the section below. The Aanusvara 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 (෴ 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 ක් 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 + ◌ෙ	කෙ	Connects to the left of consonant
ක + ◌ැ	කැ	Connects to the right of consonant
ක + ◌ි	කි	Connects to the top of consonant
ක + ◌ූ	කූ	Connects to the base of consonant
ක + ◌ෞ	කෞ	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 + ◌් + ර	ක්‍ර
ක + ◌් + ෂ	ක්‍ෂ

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

Decomopsed Form	Unicodes of Decomposed Form	Equivalent Composed Form	Unicode of Composed Form
ො ්	0DDC + 0DCA	ෝ	0DDD
ෙ ා	0DD9 + 0DCF	ො	0DDC
ෙ ෟ	0DD9 + 0DDF	ෞ	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
◌ෘ ◌ෘ = ◌ෲ	0DD8 0DD8	0DF2	1410 0020 0002	SINHALA VOWEL SIGN DIGA GAETTA-PILLA
ෙ◌ ◌් = ◌ේ	0DD9 0DCA	0DDA	141A 0020 0002	SINHALA VOWEL SIGN DIGA KOMBUVA
◌ො ◌් = ◌ෝ	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 IIYANNA
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

ඐ	0D90	1373 0020 0002	SINHALA LETTER ILUUYANNA
එ	0D91	1376 0020 0002	SINHALA LETTER EYANNA
ඒ	0D92	1379 0020 0002	SINHALA LETTER EEYANNA
ඓ	0D93	137C 0020 0002	SINHALA LETTER AIYANNA
ඔ	0D94	1380 0020 0002	SINHALA LETTER OYANNA
ඕ	0D95	1383 0020 0002	SINHALA LETTER OOYANNA
ඖ	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
ඛ	0D9B	1393 0020 0002	SINHALA LETTER MAHAAPRAANA KAYANNA
ග	0D9C	1396 0020 0002	SINHALA LETTER ALPAPRAANA GAYANNA
ඝ	0D9D	1399 0020 0002	SINHALA LETTER MAHAAPRAANA GAYANNA
ඞ	0D9E	139A 0020 0002	SINHALA LETTER KANTAJA NAASIKYAYA
ඟ	0D9F	139C 0020 0002	SINHALA LETTER SANYAKA GAYANNA
ච	0DA0	13A0 0020 0002	SINHALA LETTER ALPAPRAANA CAYANNA
ඡ	0DA1	13A3 0020 0002	SINHALA LETTER MAHAAPRAANA CAYANNA
ජ	0DA2	13A6 0020 0002	SINHALA LETTER MAHAAPRAANA JAYANNA
ඣ	0DA3	13A9 0020 0002	SINHALA LETTER MAHAAPRAANA JAYANNA
ඤ	0DA4	13AC 0020 0002	SINHALA LETTER TAALUJA NAASIKYAYA
ඥ	0DA5	13AF 0020 0002	SINHALA LETTER TAALUJA SANYOOGA NAAKSIKYAYA
ඦ	0DA6	13B0 0020 0002	SINHALA LETTER SANYAKA JAYANNA
ට	0DA7	13B3 0020 0002	SINHALA LETTER ALPAPRAANA TTAYANNA
ඨ	0DA8	13B6 0020 0002	SINHALA LETTER MAHAAPRAANA TTAYANNA
ඩ	0DA9	13B9 0020 0002	SINHALA LETTER ALPAPRAANA DDAYANNA
ඪ	0DAA	13C0 0020 0002	SINHALA LETTER MAHAAPRAAN DDAYANNA
ණ	0DAB	13C3 0020 0002	SINHALA LETTER MUURDHAJA NAYANNA
ඬ	0DAC	13C6 0020 0002	SINHALA LETTER SANYAKA DDAYANNA
ත	0DAD	13C9 0020 0002	SINHALA LETTER ALPAPRAANA TAYANNA
ථ	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
ඳ	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
ළ	0DC5	13FA 0020 0002	SINHALA LETTER MUURDHAJA LAYANNA
ෆ	0DC6	13FC 0020 0002	SINHALA LETTER FAYANNA
← Dependent Vowels →			
◌ා	0DCF	13FF 0020 0002	SINHALA VOWEL SIGN AELA-PILLA
◌ැ	0DD0	1400 0020 0002	SINHALA VOWEL SIGN KETTI AEDAPILLA
◌ෑ	0DD1	1403 0020 0002	SINHALA VOWEL SIGN DIGA AEDAPILLA
◌ි	0DD2	1406 0020 0002	SINHALA VOWEL SIGN KETTI ISPILLA
◌ී	0DD3	1409 0020 0002	SINHALA VOWEL SIGN DIGA IS-PILLA
◌ු	0DD4	140A 0020 0002	SINHALA VOWEL SIGN KETTI PAAPILLA
◌ූ	0DD6	140C 0020 0002	SINHALA VOWEL SIGN DIGA PAAPILLA
◌ෘ	0DD8	140F 0020 0002	SINHALA VOWEL SIGN GAETTAPILLA
◌ෲ	0DF2	1410 0020 0002	SINHALA VOWEL SIGN DIGA GAETTA-PILLA
◌ෘ ◌ෘ	0DD8 0DD8	1410 0020 0002	SINHALA VOWEL SIGN DIGA GAETTA-PILLA
◌ෟ	0DDF	1413 0020 0002	SINHALA VOWEL SIGN GAYANUKITTA
◌ෳ	0DF3	1416 0020 0002	SINHALA VOWEL SIGN DIGA GAYANUKITTA
◌ෙ	0DD9	1419 0020 0002	SINHALA VOWEL SIGN KOMBUVA
◌ේ	0DDA	141A 0020 0002	SINHALA VOWEL SIGN DIGA KOMBUVA
◌ෙ ◌්	0DD9 0DCA	141A 0020 0002	SINHALA VOWEL SIGN DIGA KOMBUVA
◌ෛ	0DDB	141C 0020 0002	SINHALA VOWEL SIGN KOMBU DEKA
◌ෙ ◌ෙ	0DD9 0DD9	141C 0020 0002	SINHALA VOWEL SIGN KOMBU DEKA
◌ො	0DDC	141F 0020 0002	SINHALA VOWEL SIGN KOMBUVA HAA AELA-PILLA
◌ෙ ◌ා	0DD9 0DCF	141F 0020 0002	SINHALA VOWEL SIGN KOMBUVA HAA AELA-PILLA
◌ෝ	0DDD	1420 0020 0002	SINHALA VOWEL SIGN KOMBUVA HAA DIGA AELA-PILLA
◌ො ◌්	0DDC DCA	1420 0020 0002	SINHALA VOWEL SIGN KOMBUVA HAA DIGA AELA-PILLA

ෙ◌ ◌ා ◌්	0DD9 0DCF DCA	1420 0020 0002	SINHALA VOWEL SIGN KOMBUVA HAA DIGA AELA-PILLA
ෙ◌ෟ	0DDE	1423 0020 0002	SINHALA VOWEL SIGN KOMBUVA HA GAYANUKITTA
ෙ◌ ◌ෟ	0DD9 0DDF	1423 0020 0002	SINHALA VOWEL SIGN KOMBUVA HA GAYANUKITTA
◌්	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	ආපසුඵකවා	කාරකයා
කටුවා	කංසාරිකි	ආපසුයකවා	කැ
උක්කණ	කඃසක	ආපසුයාම	කැරළිකාර
ආපසුයකවා	කකණටකයා	ඇළය	කැරළිකාරක
ඇඳපු	කකළ	ඇඳපු	කැරළිහසකවා
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	කඃසක	කු
කරාබු	කැ	කකණටකයා	කුබුදිකවා
කෘමිල	කෑ	කකළ	කුබුදු
අකක්	කි	කකා	කුබුද්දකවා
කොණ්ඔ	කෘ	කක්වයි	කූ
කැරළිකාරක	කො	කටු	කෘ
කැරළිහසකවා	කෞ	කටුක	කෘග
කිකමිද	ක්	කටුව	කෘමිල
කෙක්ක	කෲ	කටුවා	කෲ
කිකුටු	කෟ	කරරුස	කෟ
කල්කියාව	කී	කරවක	කෟ

කුබුද්දකවා	කු	කරස	කෙ
කිකි	කූ	කරාබු	කෙකි
කාරකයා	කෟ	කල්කණඩු	කෙක්ක
කාරක	කෙ	කල්කියාව	කේ
කෘග	කේ	කවබැඳුම	කෛ
කෙකි	කෛ	කවාරම	කො
කැරළිකාර	අකණය	කවිකය	කොණ්ඹ
ආපසුඵකවා	කරරුස	කවු	කෞ
ඊගස	කුබුදු	කවෙ	ක්

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.

9. Tamil

Tamil is a Southern Dravidian language [51]. It is currently spoken by about 77 million people around the world with 68 million speakers residing in India mostly in the state of Tamil Nadu. It is one the official language in India, Sri Lanka and Singapore.

Tamil language is written in Tamil script which descends from South Brahmi script and dates back to 500 BC [8, 52]. It is a syllabic writing system, like other Indic systems, written without a top-line characteristic of South Brahmi scripts and different from the North Brahmi scripts.

9.1. Writing System

9.1.1. Character Set

Tamil has fewer characters, a total of 18 consonants, 12 independent vowels and 11 dependent vowels (schwa, the twelfth vowel, is implied with each consonant and thus not written explicitly). These are given in Figures 9.1 and 9.2.

அ ஆ இ ஈ உ ஊ எ ஏ ஐ ஒ ஓ ஔ

Independent Vowels

◌ா ◌ி ◌ீ ◌ு ◌ூ ◌ெ ◌ே ◌ை ◌ொ ◌ோ ◌ௌ

Dependent Vowels

Figure 9.1. Tamil Vowels

க ங ச ஞ ட ண த ந ப ம ய ர ல வ ழ ள ற ன

Figure 9.2. Tamil Consonants

Tamil borrows five special consonants to represent Sanskrit loan words. These are known as Grantha characters [52] and are shown below.

ஜ ஸ ஷ ஹ க்ஷ

Figure 9.3. Additional Tamil Consonants (For Loan Words)

Tamil also has two special characters, Virama and Aytam. Virama is used to cancel the implicit vowel with each consonant. Aytam causes spirintization, turning [p] into [f] and [j] into [z]. Their use is shown in Table 9.1.

Table 9.1. Virama and Aytam Characters in Tamil

Name	Glyph	Usage
Virma	◌்	ப்
Aytam	◌ஃ	பஃ

Tamil has its own set of numerals but these are rarely used, and normally 0, 1, 2… 9 are used.

௦ ௧ ௨ ௩ ௪ ௫ ௬ ௭ ௮ ௯

Figure 9.4. Tamil Numerals

In addition, Tamil has special characters as multipliers for 10, 100 and 1000, shown in Figure 9.5. Thus, ௩, ௰௩, ௩௰, ௩௰௩ represent 3, 13, 30 and 33 respectively [5].

௰ ௱ ௲

Figure 9.5. Tamil Multipliers for 10, 100, and 1000

Moreover, there are some special symbols for day, month, year, debit, credit, as above, rupee and numeral in Tamil shown in Figure 9.6 below [4].

௳ ௴ ௵ ௶ ௷ ௸ ௹ ௺

Figure 9.6. Special Signs

9.1.2. Script Details

Tamil is written from left to right. Letters are uncased and are grouped based on their place and manner of articulation, like other Indic scripts.

9.1.2.1. Consonants and Vowels

Tamil has a syllabic writing system. Vowels and consonants are not represented as an individual unit, rather as syllabic units in which consonant has an inherent [a] vowel, if not otherwise specified. For example Tamil Letter KA க has [ka] 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, Virama is placed at top of the consonant to cancel the [a] sound. So க் 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 9.2 below shows these cases.

Table 9.2. Dependent Vowels with the Consonant [h]

Consonant + Dependent Vowel	Joined Form	Comment
ஹ+ ◌ெ	ஹெ	Connects to the left of consonant
ஹ + ◌ா	ஹா	Connects to the right of consonant
ஹ + ◌ீ	ஹீ	Connects to the top of consonant
ஹ + ◌ு	ஹு	Connects to the base of consonant
ஹ + ◌ௌ	ஹௌ	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.

9.1.2.2. Conjunct Consonants and Consonantal Vowel Ligatures

Tamil, unlike other Indic based languages do not form conjunct consonants except for the case of letter KSA which is formed as letter KA + Virama + SA (க்+ஷ = க்ஷ). However, Tamil frequently forms consonant-vowel ligatures. Same vowels might form variety of different shapes when combined with different consonants. This is shown in the figure below as different consonants combine with the long vowel ◌ூ. For a more comprehensive list, see [52].

க + ◌ூ = கூ
ச + ◌ூ = சூ
த + ◌ூ = தூ
ப + ◌ூ = பூ
ம + ◌ூ = மூ

Figure 9.7. Consonant Vowel Ligatures in Tamil

9.2. Collation

This section elaborates on this collation sequence for Tamil and an algorithmic implementation using UCA [2].

All characters have primary level significance for collation purposes. Numerals and currency symbols are given smallest weight. These are followed by modifiers, independent vowels, consonants and finally dependent vowels. However, before collation can be applied, some text processing is required. These details are also given below.

9.2.1. Text Processing

9.2.1.1. Reordering

As shown in Table 9.1 above, independent vowels combine with consonants in different ways. In hand-written orthography, old type-writers and early proprietary Tamil 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 based on Unicode follow the logical order of typing. However, for the legacy fonts, re-ordering will be required before string comparisons can be performed for collation.

9.2.1.2. Virama

Virama is implicitly an integral part of most Indic scripts; however, its explicit use is sometimes optional. In case Virama is not written, a native speaker can still use the knowledge of the language to correctly recognize and pronounce the words. However, it is much more consistently used in Tamil. Consonants with Virama are lighter than the same consonant without it. This is not possible to do if separate collation elements are assigned to consonants and Virama, as per Unicode collation algorithm. A solution is to define a contraction corresponding to each consonant with a collation element with lesser value compared to the consonant without the Virama, as given in the collation table later.

9.2.1.3. Normalization

Many Tamil 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 9.3 below.

Table 9.3. Normalization in Tamil

Decomposed Form	Unicodes of Decomposed Form	Equivalent Composed Form	Unicode of Composed Form

ெ◌ + ◌ா	0BC6 + 0BBE	ெ◌ா	0BCA
ே◌ + ◌ா	0BC7 + 0BBE	ே◌ா	0BCB
ெ◌+ ◌ௗ	0BC6+ 0BD7	ெ◌ௗ	0BCC
ஒ + ◌ௗ	0B92 + 0BD7	ஔ	0B94

9.2.1.4. Contraction

In Tamil, a few sequences of encoded characters map onto contracted linguistic units, which have distinct role in collation, different from their constituents. These contractions need to be assigned appropriate collation elements. These include the KSA character and the consonants with Virama (as discussed earlier). Examples for these two cases are given in Table 9.4 below.

Table 9.4. Contraction to Single Collation Element from Multiple Unicodes

Glyph	Unicodes of Decomposed Form	Composed Form	Collation Element	Name
க + ◌் + ஷ	0B95 + 0BCD + 0BB7	க்ஷ	13B6 0020 0002	TAMIL LETTER KSA
ம + ◌்	0BAE + 0BCD	ம்	1392 0020 0002	TAMIL LETTER M

9.2.1.5. Consonantal Vowel Ligatures

As discussed, in some cases when vowels combine with consonants, they form a conjoined shape which is different from simple concatenation. The formation of these consonantal vowel ligatures is a visual phenomenon and does not change the encoding or the linguistic entities in any complex way. Thus, it is not relevant for collation process.

9.2.2. Unicode Collation Elements

In order to realize Tamil collation, 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 [53, 54].

Table 9.5. Tamil Collation Elements

Glyph	Unicode	Collation Elements	Unicode Name
← Numerals and Various Signs →			
௦	0BE6	0E29 0020 0002	TAMIL DIGIT ZERO
௧	0BE7	0E2A 0020 0002	TAMIL DIGIT ONE

௨	0BE8	0E2B 0020 0002	TAMIL DIGIT TWO
௩	0BE9	0E2C 0020 0002	TAMIL DIGIT THREE
௪	0BEA	0E2D 0020 0002	TAMIL DIGIT FOUR
௫	0BEB	0E2E 0020 0002	TAMIL DIGIT FIVE
௬	0BEC	0E2F 0020 0002	TAMIL DIGIT SIX
௭	0BED	0E30 0020 0002	TAMIL DIGIT SEVEN
௮	0BEE	0E31 0020 0002	TAMIL DIGIT EIGHT
௯	0BEF	0E32 0020 0002	TAMIL DIGIT NINE
௰	0BF0	0E33 0020 0002	TAMIL NUMBER TEN
௱	0BF1	0E34 0020 0002	TAMIL NUMBER ONE HUNDERED
௲	0BF2	0E35 0020 0002	TAMIL NUMBER ONE THOUSAND
௳	0BF3	0E36 0020 0002	TAMIL DAY SIGN
௴	0BF4	0E37 0020 0002	TAMIL MONTH SIGN
௵	0BF5	0E38 0020 0002	TAMIL YEAR SIGN
௶	0BF6	0E39 0020 0002	TAMIL DEBIT SIGN
௷	0BF7	0E3A 0020 0002	TAMIL CREDIT SIGN
௸	0BF8	0E3B 0020 0002	TAMIL AS ABOVE SIGN
௹	0BF9	0E3C 0020 0002	TAMIL RUPEE SIGN
௺	0BFA	0E3E 0020 0002	TAMIL NUMBER SIGN
◌்	0BCD	1350 0020 0002	TAMIL SIGN VIRMA
◌ஃ	0B83	1353 0020 0002	TAMIL SIGN VISARGA
←Independent Vowels Primary Level→			
அ	0B85	1356 0020 0002	TAMIL LETTER A
ஆ	0B86	1359 0020 0002	TAMIL LETTER AA
இ	0B87	135C 0020 0002	TAMIL LETTER I
ஈ	0B88	135F 0020 0002	TAMIL LETTER II
உ	0B89	1360 0020 0002	TAMIL LETTER U
ஊ	0B8A	1363 0020 0002	TAMIL LETTER UU
எ	0B8E	1366 0020 0002	TAMIL LETTER E
ஏ	0B8F	1369 0020 0002	TAMIL LETTER EE
ஐ	0B90	136C 0020 0002	TAMIL LETTER AI
ஒ	0B92	136F 0020 0002	TAMIL LETTER O
ஓ	0B93	1370 0020 0002	TAMIL LETTER OO
ஔ	0B94	1373 0020 0002	TAMIL LETTER AU
ஒ ◌ௗ	0B92 0BD7	1373 0020 0002	TAMIL LETTER AU
←Consonants→			
க ◌்	0B95 0BCD	1375 0020 0002	TAMIL LETTER K
க	0B95	1376 0020 0002	TAMIL LETTER KA

ங ◌்	0B99 0BCD	1378 0020 0002	TAMIL LETTER NG
ங	0B99	1379 0020 0002	TAMIL LETTER NGA
ச ◌்	0B9A 0BCD	137B 0020 0002	TAMIL LETTER C
ச	0B9A	137C 0020 0002	TAMIL LETTER CA
ஞ ◌்	0B9E 0BCD	137F 0020 0002	TAMIL LETTER NY
ஞ	0B9E	1380 0020 0002	TAMIL LETTER NYA
ட ◌்	0B9F 0BCD	1382 0020 0002	TAMIL LETTER TT
ட	0B9F	1383 0020 0002	TAMIL LETTER TTA
ண ◌்	0BA3 0BCD	1385 0020 0002	TAMIL LETTER NN
ண	0BA3	1386 0020 0002	TAMIL LETTER NNA
த ◌்	0BA4 0BCD	1388 0020 0002	TAMIL LETTER T
த	0BA4	1389 0020 0002	TAMIL LETTER TA
ந ◌்	0BA8 0BCD	138B 0020 0002	TAMIL LETTER N
ந	0BA8	138C 0020 0002	TAMIL LETTER NA
ப ◌்	0BAA 0BCD	138F 0020 0002	TAMIL LETTER P
ப	0BAA	1390 0020 0002	TAMIL LETTER PA
ம ◌்	0BAE 0BCD	1392 0020 0002	TAMIL LETTER M
ம	0BAE	1393 0020 0002	TAMIL LETTER MA
ய ◌்	0BAF 0BCD	1395 0020 0002	TAMIL LETTER Y
ய	0BAF	1396 0020 0002	TAMIL LETTER YA
ர ◌்	0BB0 0BCD	1398 0020 0002	TAMIL LETTER R
ர	0BB0	1399 0020 0002	TAMIL LETTER RA
ல ◌்	0BB2 0BCD	139A 0020 0002	TAMIL LETTER L
ல	0BB2	139B 0020 0002	TAMIL LETTER LA
வ ◌்	0BB5 0BCD	139C 0020 0002	TAMIL LETTER V
வ	0BB5	139D 0020 0002	TAMIL LETTER VA
ழ ◌்	0BB4 0BCD	139F 0020 0002	TAMIL LETTER LLL
ழ	0BB4	13A0 0020 0002	TAMIL LETTER LLLA
ள ◌்	0BB3 0BCD	13A2 0020 0002	TAMIL LETTER LL
ள	0BB3	13A3 0020 0002	TAMIL LETTER LLA
ற ◌்	0BB1 0BCD	13A5 0020 0002	TAMIL LETTER RR
ற	0BB1	13A6 0020 0002	TAMIL LETTER RRA
ன ◌்	0BA9 0BCD	13A8 0020 0002	TAMIL LETTER NNN
ன	0BA9	13A9 0020 0002	TAMIL LETTER NNNA
ஜ ◌்	0B9C 0BCD	13AB 0020 0002	TAMIL LETTER J
ஜ	0B9C	13AC 0020 0002	TAMIL LETTER JA
ஸ ◌்	0BB8 0BCD	13AE 0020 0002	TAMIL LETTER S

ஸ	0BB8	13AF 0020 0002	TAMIL LETTER SA
ஷ ◌்	0BB7 0BCD	13B0 0020 0002	TAMIL LETTER SS
ஷ	0BB7	13B1 0020 0002	TAMIL LETTER SSA
ஹ ◌்	0BB9 0BCD	13B2 0020 0002	TAMIL LETTER H
ஹ	0BB9	13B3 0020 0002	TAMIL LETTER HA
க ◌் ஷ ◌்	0B95 0BCD 0BB7 0BCD	13B5 0020 0002	TAMIL LETTER KS
க ◌் ஷ	0B95 BCD 0BB7	13B6 0020 0002	TAMIL LETTER KSA
ஶ்	0BB6 0BCD	13B8 0020 0002	TAMIL LETTER SH
ஶ	0BB6	13B9 0020 0002	TAMIL LETTER SHA
← Dependent Vowels →			
◌ா	0BBE	13C0 0020 0002	TAMIL VOWEL SIGN AA
◌ி	0BBF	13C3 0020 0002	TAMIL VOWEL SIGN I
◌ீ	0BC0	13C6 0020 0002	TAMIL VOWEL SIGN II
◌ு	0BC1	13C9 0020 0002	TAMIL VOWEL SIGN U
◌ூ	0BC2	13CA 0020 0002	TAMIL VOWEL SIGN UU
◌ெ	0BC6	13CC 0020 0002	TAMIL VOWEL SIGN E
◌ே	0BC7	13D0 0020 0002	TAMIL VOWEL SIGN EE
◌ை	0BC8	13D3 0020 0002	TAMIL VOWEL SIGN AI
◌ொ	0BCA	13D6 0020 0002	TAMIL VOWEL SIGN O
◌ெ ◌ா	0BC6 0BBE	13D6 0020 0002	TAMIL VOWEL SIGN O
◌ோ	0BCB	13D9 0020 0002	TAMIL VOWEL SIGN OO
◌ே ◌ா	0BC7 0BBE	13D9 0020 0002	TAMIL VOWEL SIGN OO
◌ௌ	0BCC	13DC 0020 0002	TAMIL VOWEL SIGN AU
◌ெ ◌ள	0BC6 0BD7	13DC 0020 0002	TAMIL VOWEL SIGN AU
◌ள	0BD7	13DF 0020 0002	TAMIL AU LETTER MARK

9.2.3. Results

The collation elements were applied to sort a random set of strings of Tamil. The input and corresponding output is given in Table 9.6.

Table 9.6. Input and Corresponding Sorted Output for Tamil

Input		Output	
வெள	ஒமம்	அக்கறை	கேழல்
வௌ	ஒஷ	அககுள்	கைராசி
வௌவு	ஜயர்	அககுள	கைராட்டு

யக்தம்	ஐயர	ஆன்	கொத்து
யக்தி	எஃகு	இன்	கொத்து
யகம்	கூட்டம்	ஈசல்	கொத்தூ
முந்து	கெடு	உனற	கோரி
முந்தை	கெடுதி	ஊழ்	கோரி
பிரமன்	கேழ்	எஃகு	கோல்
பிரமை	கேழல்	எற்று	கௌ
நீக்கம்	கைராசி	ஏர்	கௌ
நீக்கல்	கைராட்டு	ஐயர்	செட்டி
நீங்கு	கொத்து	ஐயர	செட்டு
தூற்று	கொத்தூ	ஒமம்	செடி
தூறல்	கொத்து	ஒஷ	ஞாலம்
தூறு	கோரி	ஔவ	ஞாழல்
ஞாலம்	கோரி	ஔவை	தூற்று
ஞாழல்	கோல்	கஃசு	தூறல்
செட்டி	கௌ	கக்கம	தூறு
செட்டு	கௌ	கக	நீக்கம்
செடி	ஔவை	காந்தல்	நீக்கல்
கஃசு	ஔவ	காந்தள்	நீங்கு
கக்கம	ஏர்	கிங்கரன்	பிரமன்
கக	எற்று	கிங்கிணி	பிரமை
காந்தல்	ஊழ்	குலவு	முந்து
காந்தள்	உனற	குலாலன்	முந்தை
கிங்கரன்	ஈசல்	குலாவு	யக்தம்
கிங்கிணி	இன்	கூட்டடி	யக்தி
குலவு	ஆன்	கூட்டம்	யகம்
குலாலன்	அக்கறை	கெடு	வௌ
குலாவு	அககுள்	கெடுதி	வௌ
கூட்டடி	அககுள	கேழ்	வௌவு

9.3. Conclusion

Tamil 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 contract 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.

10. Urdu

Urdu derives from Indo-Aryan family of languages and shares basic linguistic structure with Hindi, the two languages being mutually understandable. However, unlike Hindi, Urdu derives more of its vocabulary from Persian and Arabic. Urdu has 104 million speakers in Pakistan, Afghanistan, India, Bangladesh and many other countries [56]. Urdu is the national language of Pakistan and a state language of India. Urdu is written using Arabic script. Perso-Arabic Nastalique style is widely used for Urdu orthography [57, 58].

10.1. Writing System

10.1.1. Character Set

Urdu uses characters from the extended Arabic character set used for Persian. It further extends this set to represent sounds which are present in Urdu but not it Arabic or Persian, including aspirated stop and alveolar consonants, and long vowels [59]. Altogether there are 58 letters in Urdu, given in Figure 10.1 ([60]; other sources may give slightly different set).

ا آ ب بھ پ پھ ت تھ ٹ ٹھ ث ج جھ چ چھ ح خ د دھ ڈ ڈھ ذ

ر رھ ڑ ڑھ ز ژ س ش ص ض ط ظ ع غ ف ق ک کھ گ گھ ل

لھ م مھ ن نھ ں نھ و وھ ہ ۃ ء ی یھ ے

Figure 10.1. Urdu Character Set

Arabic script uses letters to represent consonants. Diacritics are used to specify the vowels. Urdu has both long and short vowels. Short vowels are indicated by placing diacritics with the consonant which precedes it in the syllable. Long vowels are indicated by a combination of the diacritic on a consonant followed by an additional letter (see [59] for a detailed discussion). These diacritics (also known as Aerab) are normally not written, though are implicitly present, and thus are optional in their usage. In addition to the Aerab which specify the vowels, diacritics are also used to add consonantal sounds, e.g. for germination (i.e. duplication of consonants). These Aerab are given in Figure 10.2 with the letter ب.

بَ بِ بُ بٗ بً بٰ بّ

Figure 10.2. Urdu Diacritics

As is evident from the figure, different diacritics can occur above or below the consonant. A consonant may take two diacritics, one consonantal and the other vocalic. In case both diacritics are above the consonant, the consonantal diacritics stack under the vocalic one. These diacritics are always keyed in after the anchoring base letter.

Urdu also has honorific marks which are used to show respect, and are used with proper names. These honorifics are shown in Figure 10.3.

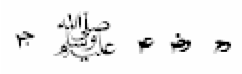

Figure 10.3. Honorific Marks in Urdu

Urdu has its own set of numerals based on numerals used in Arabic and Persian, but some numerals are unique in their shape. These numerals are listed in Figure 10.4.

۹ ۸ ۷ ٦ ۵ ۴ ۳ ۲ ۱ ۰

Figure 10.4. Urdu Numerals

10.1.2. Bidirectionality

Urdu inherits the bidirectional property from Arabic script. Urdu words are written from right to left but numbers are written from right to left, as shown in Figure 10.5. However, bidirectionality is handled at rendering level and key press sequence for Urdu alphanumeric input is same as it would be for any other uni-directional language. Thus bidirectionality has no implication on collation.

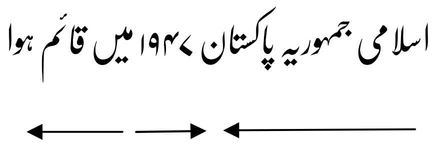

Figure 10.5. Bidirectional Urdu Text
(Arrows indicate reading direction)

10.1.3. 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. In Nastalique, the character may also change shape depending on the other characters around it. Thus, depending on the context, a single letter may take as many as 25 shapes. Cursiveness is shown in Figure 10.6.

ب ا د ش ا ہ ی   م س ج د

Spelling

بادشاہی مسجد

Cursively Written Form

Figure 10.6. Spelt-out and Cursive Version of Sample Text of Urdu in Nastalique Script

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 many shapes but its internal encoding is always U+0628. Therefore, these properties have no implication on collation.

10.2. Collation

Urdu collation sequence has been standardized and published by National Language Authority for Pakistan [60]. 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.

10.2.1. Text Processing

10.2.1.1. Inconsistent Use of Space

Nastalique style of writing does not have the concept of space to separate words. Similar to South-East Asian scripts like Lao, Thai and Khmer, Urdu readers are expected to parse the ligatures into words as they read along the text. In typing, space is used to get the right character shapes. To achieve this end, it is sometimes used within a word to break the word into constituent ligatures, as in word یونی ورسٹی. However, if the ligature form is achieved without the use of space, it is sometimes not even used in between two words, e.g. اردوخط is visually correct sequence of two words for the readers but has no space between them. 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 Urdu and therefore the input for collation must be manually cleaned.

10.2.1.2. Diacritics for Loan Words

The diacritics used for Urdu are given in the figure above. However, there are additional diacritics which are sometimes used with loan words from Arabic. Though not part of Urdu, they have to be processed in case of loan words. Thus, they are also included in the collation element table.

10.2.1.3. Normalization

Two kinds of normalization are required for Urdu. 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 ی may be represented by Unicode points U+0649, U+064A, and U+06CC in Urdu[1]. 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. This category includes two sub-categories. One category combines marks and base characters to form other characters. Other combines multiple base characters to form a ligature. Table 10.1 below gives some examples.

[1] These codes are normally used in Urdu corpus online to represent ی character. Additional codes in Arabic Presentations Forms are not listed here. Unicode doe s not recommend the use of this area, which was originally used for backward compatibility with legacy systems.

Table 10.1. Composed and Decomposed Forms of an Urdu Character and a Ligature

Glyph	Unicode	Individual letters/marks	Unicode Points
آ	0622	ا ٓ	0653 0627
لا	FEFB	ل ا	0627 06F1

There are many such characters and 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 Urdu language (U+0660 – U+0669), where a separate similar looking set is also encoded (U+06F0 – U+06F9) for use in Arabic language.

10.2.1.4. Contraction

In Urdu character ھ (U+06BE or U+0647[2]) combines with most obstruents[3] to represent their aspirated version. 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 10.7.

ب + ھ = بھ

ج + ھ = جھ

ڈ+ھ = ڈھ

Figure 10.7. Contraction of Letters with ھ in Urdu

There is no Unicode point available to directly encode the contracted form for aspirated obstruents.

[2] Not recommended for use in Urdu but is used in the online Urdu corpus.

[3] Phonological term for all sounds which cause a constriction in the oral tract during articulation.

10.2.2. Unicode Collation Elements

Collation Elements for Urdu character set are given in Table 10.2 below.

Table 10.2. Urdu Collation Elements

Glyph	Unicode	Collation Elements	Unicode Name
ZWNJ	200C	0000 0010 0002	ZERO WIDTH NON-JOINER
← Diacrtics→			
◌ْ	0652	0000 00C4 0002	ARABIC SUKUN
◌َ	064E	0000 00C9 0002	ARABIC FATHA
◌ِ	0650	0000 00CA 0002	ARABIC KASRA
◌ُ	064F	0000 00CB 0002	ARABIC DAMMA
◌ٰ	0670	0000 00CD 0002	ARABIC LETTER SUPERSCRIPT ALEF
◌ٖ	0656	0000 00D5 0002	ARABIC SUBSCRIPT ALEF
◌ٗ	0657	0000 00D8 0002	ARABIC INVERTED DAMMA
◌ً	064B	0000 00DB 0002	ARABIC FATHATAN
◌ٍ	064D	0000 00DE 0002	ARABIC KASRATAN
◌ٌ	064C	0000 00E2 0002	ARABIC DAMMATAN
◌ٔ	0654	0000 00E5 0002	ARABIC HAMZA ABOVE
◌ّ	0651	0000 00E8 0002	ARABIC SHADDA
◌٘	0658	0000 00EA 0002	ARABIC MARK NOON GHUNNA
◌ٓ	0653	0000 00F1 0002	ARABIC MADDAH ABOVE
← 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
◌ؔ	0614	0000 0000 004A	ARABIC SIGN TAKHALLUS
← Punctuation Marks→			
؀	0600	0000 0000 0000	ARABIC NUMBER SIGN
؁	0601	0000 0000 0000	ARABIC SIGN SANAH
؂	0602	0000 0000 0000	ARABIC FOOTNOTE MARKER
؃	0603	0000 0000 0000	ARABIC SIGN SAFHA
◌ؕ	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
؎	060E	0000 0000 0000	ARABIC POETIC VERSE SIGN
؏	060F	0000 0000 0000	ARABIC SIGN MISRA
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

ؤ	0624	[13BD 0020 0002],[0000 00E5 0002]	ARABIC LETTER WAW WITH HAMZA ABOVE
ئ	0626	[13C9 0020 0002],[0000 00E5 0002]	ARABIC LETTER CHOTI YEH WITH HAMZA ABOVE
أ	0623	[1350 0020 0002],[0000 00E5 0002]	ARABIC LETTER ALEF WITH HAMZA ABOVE
← Numerals →			
۰	06F0	0E29 0020 0002	ARABIC-INDIC DIGIT ZERO
۱	06F1	0E2A 0020 0002	ARABIC-INDIC DIGIT ONE
۲	06F2	0E2B 0020 0002	ARABIC-INDIC DIGIT TWO
۳	06F3	0E2C 0020 0002	ARABIC-INDIC DIGIT THREE
۴	06F4	0E2D 0020 0002	ARABIC-INDIC DIGIT FOUR
۵	06F5	0E2E 0020 0002	ARABIC-INDIC DIGIT FIVE
۶	06F6	0E2F 0020 0002	ARABIC-INDIC DIGIT SIX
۷	06F7	0E30 0020 0002	ARABIC-INDIC DIGIT SEVEN
۸	06F8	0E31 0020 0002	ARABIC-INDIC DIGIT EIGHT
۹	06F9	0E32 0020 0002	ARABIC-INDIC DIGIT NINE
ا	0627	1350 0020 0002	ARABIC LETTER ALEF
آ	0627 0653	1351 0020 0002	ARABIC LETTER ALEF WITH MADDA ABOVE
آ	0622	1351 0020 0002	ARABIC LETTER ALEF WITH MADDA ABOVE
ب	0628	1352 0020 0002	ARABIC LETTER BEH
بھ	0628 06BE	1353 0020 0002	ARABIC LETTER BEH + ARABIC LETTER HEH DOCHASHMEE

پ	067E	1354 0020 0002	ARABIC LETTER PEH
پھ	067E 06BE	1355 0020 0002	ARABIC LETTER PEH + ARABIC LETTER HEH DOCHASHMEE
ت	062A	1357 0020 0002	ARABIC LETTER TEH
تھ	062A 06BE	1358 0020 0002	ARABIC LETTER THE + ARABIC LETTER HEH DOCHASHMEE
ٹ	0679	135A 0020 0002	ARABIC LETTER TTEH
ٹھ	0679 06BE	135B 0020 0002	ARABIC LETTER TTEH + ARABIC LETTER HEH DOCHASHMEE
ث	062B	135D 0020 0002	ARABIC LETTER THEH
ج	062C	135E 0020 0002	ARABIC LETTER JEEM
جھ	062C 06BE	135F 0020 0002	ARABIC LETTER JEEM + ARABIC LETTER HEH DOCHASHMEE
چ	0686	1361 0020 0002	ARABIC LETTER TCHEH
چھ	0686 06BE	1362 0020 0002	ARABIC LETTER TCHEH + ARABIC LETTER HEH DOCHASHMEE
ح	062D	1364 0020 0002	ARABIC LETTER HAH
خ	062E	1365 0020 0002	ARABIC LETTER KHAH
د	062F	1369 0020 0002	ARABIC LETTER DAL
دھ	062F 06BE	136A 0020 0002	ARABIC LETTER DAL + ARABIC LETTER HEH DOCHASHMEE
ڈ	0688	136B 0020 0002	ARABIC LETTER DDAL
ڈھ	0688 06BE	136C 0020 0002	ARABIC LETTER DDAL + ARABIC LETTER HEH DOCHASHMEE
ذ	0630	1370 0020 0002	ARABIC LETTER THAL
ر	0631	1375 0020 0002	ARABIC LETTER REH

رھ	0631 06BE	1376 0020 0002	ARABIC LETTER REH + ARABIC LETTER HEH DOCHASHMEE
ڑ	0691	1377 0020 0002	ARABIC LETTER RREH
ڑھ	0691 06BE	1378 0020 0002	ARABIC LETTER RREH + ARABIC LETTER HEH DOCHASHMEE
ز	0632	137C 0020 0002	ARABIC LETTER ZAIN
ژ	0698	137E 0020 0002	ARABIC LETTER JEH
س	0633	1381 0020 0002	ARABIC LETTER SEEN
ش	0634	1382 0020 0002	ARABIC LETTER SHEEN
ص	0635	1387 0020 0002	ARABIC LETTER SAD
ض	0636	1388 0020 0002	ARABIC LETTER DAD
ط	0637	138C 0020 0002	ARABIC LETTER TAH
ظ	0638	138D 0020 0002	ARABIC LETTER ZAH
ع	0639	138F 0020 0002	ARABIC LETTER AIN
غ	063A	1390 0020 0002	ARABIC LETTER GHAIN
ف	0641	1393 0020 0002	ARABIC LETTER FEH
ق	0642	139B 0020 0002	ARABIC LETTER QAF
ک	06A9	139F 0020 0002	ARABIC LETTER KEHEH
کھ	06A9 06BE	13A2 0020 0002	ARABIC LETTER KEHEH + ARABIC LETTER HEH DOCHASHMEE
گ	06AF	13A5 0020 0002	ARABIC LETTER GAF
گھ	06AF 06BE	13A6 0020 0002	ARABIC LETTER GAF + ARABIC LETTER HEH DOCHASHMEE

ل	0644	13AB 0020 0002	ARABIC LETTER LAM
لھ	0644 06BE	13AC 0020 0002	ARABIC LETTER LAM + ARABIC LETTER HEH DOCHASHMEE
م	0645	13B0 0020 0002	ARABIC LETTER MEEM
مھ	0645 06BE	13B1 0020 0002	ARABIC LETTER MEEM + ARABIC LETTER HEH DOCHASHMEE
ن	0646	13B4 0020 0002	ARABIC LETTER NOON
نھ	0646 06BE	13B5 0020 0002	ARABIC LETTER NOON + ARABIC LETTER HEH DOCHASHMEE
ں	06BA	13B9 0020 0002	ARABIC LETTER NOON GHUNNA
ںھ	06BA 06BE	13BA 0020 0002	ARABIC LETTER NOON GHUNNA + ARABIC LETTER HEH DOCHASHMEE
و	0648	13BD 0020 0002	ARABIC LETTER WAW
وھ	0648 06BE	13BE 0020 0002	ARABIC LETTER WAW + ARABIC LETTER HEH DOCHASHMEE
ہ	06C1	13C2 0020 0002	ARABIC LETTER HEH GOAL
ھ	06BE	13C4 0020 0002	ARABIC LETTER HEH DOCHASHMEE
ۃ	06C3	13C6 0020 0002	ARABIC LETTER TEH MARBUTA GOAL
ء	0621	13C7 0020 0002	ARABIC LETTER HAMZA
ی	06CC	13C9 0020 0002	ARABIC LETTER FARSI YEH
یھ	06CC06BE	13CB 0020 0002	ARABIC LETTER FARSI YEH + ARABIC LETTER HEH DOCHASHMEE
ے	06D2	13CE 0020 0002	ARABIC LETTER YEH BARREE

10.3. Results

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

Table 10.3. Input and Corresponding Sorted Output for Urdu

Sample Output		Sample Input	
دائرة	اب	بہن	بھنگی
دائرةالمعروف	ابھی	بی بی	اگنا
زکوت	اگنا	عمُر	بیٹی
زکوه	ایمان	دائرةالمعروف	دائرہ
زکوة	آب	آبن	گَنَّا
زکوٰة	آبن	عمر	عمر
عمر	بَن	ماں	گَنَا
عمر	بِن	بی	آب
عمُر	بُن	گَنَّا	ابھی
عُمر	بہن	زکوٰة	ایمان
گَنا	بی	بے	ٹیلیفون
گَنَا	بیبی	زکوة	عُمر
گَنّا	بیٹی	زکوت	مان
گَنَّا	بے	زکوه	ٹیلی‌فون
گَنَّا	بھنگی	بِن	گَنا
مان	ٹیلیفون	دائرة	اب
ماں	ٹیلیفون	بُن	گَنّا
	دائرہ		بَن

10.4. Conclusion

Sorting in Urdu 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 Urdu collation. However, regular sorting algorithm is applicable after appropriate text processing is done and collation elements are assigned.

11. Discussion and Conclusion

The current work addresses a variety of cases for development of collation sequences across languages. As has been shown, collation is a complex linguistic phenomenon dependent on a variety of factors deriving mostly from the writing and speaking system of a language. Computing adds another layer of complexity to collation, as the sorting process is further dependent by the encoding system. Thus, both linguistic and computing rigor is required to find a solution which is conventionally acceptable by speakers of the language. In certain cases, multiple collations are also culturally acceptable[1]. In such cases, all the collations must be separately implemented and the choice of collation should be left to the user and context.

Most languages give a varying degree of importance to their character repository. Core characters can take primary level importance. Some other characters may take secondary importance, and so on, until the other end of the spectrum, where language may also have characters which are ignorable for collation purpose. Asian scripts and languages, like other scripts, also have a variety of levels of collation, motivated by different factors.

Marks and symbols are also an important part of orthography which may have significant influence on collation. Like other languages and scripts, Asian writing systems and languages employ them for a variety of reasons. Marks specify and/or modify the consonantal, vocalic and other phonological material in words. Lao uses marks to specify tones. Lao, Dzongkha, Sindhi, Urdu and other languages use marks to specify vowels. Bengali uses a mark to suppress inherent vowel to form closed syllables. Other phonological properties like germination (e.g. in Urdu), nasalization (e.g. in Bengali), aspiration (e.g. in Dzongkha) and spirintization (e.g. in Tamil) are also indicated by marks. Marks and symbols are also used to specify information at higher linguistic levels, e.g. to specify syllable boundary (e.g. in Dzongkha) or phrase boundary (e.g. in Sinhala and Dzongkha), or to mark levels of respect and honor (e.g. in Urdu). And, as anticipated, these marks affect collation to a varying degree. Some marks have primary level collation weight, while others are ignorable for collation. A level of complexity is also added due to the fact that some of these marks are optionally used, e.g. the Virama in Bengali and the Aerab in Urdu. When they are not written, they are re-constructed by a human reader intelligently, even when being collated. However, this is difficult to model.

[1] Such multiplicity is also observed in other cultural conventions, e.g. many cultures have multiple calendars.

Character ordering, and thus collation, is more strongly motivated by script than language. For example, Hindi is very similar to Urdu as a language[2], the former written with Devanagari script (similar to Bengali script discussed in this volume) and latter with Arabic script. Hindi characters are ordered by sound, with consonants grouped according to place of articulation, starting from velars all to way to labials. However, Urdu characters are grouped according to shape, as have been done in Arabic script based languages, and not according to their place of articulation. Sindhi shows same behavior, even though it significantly extends the basic Arabic script. Furthermore, Hindi also has independent and dependent vowels which show context sensitive collation and orthography (like Bengali), much different from Urdu.

As has been observed for the languages discussed, a key factor which influences the collation weight of a character is its context, latter sometimes also marked with change in orthography. Dzongkha and Laos show significant change in consonant behavior depending on where it occurs in a syllable (which is eventually determined by which characters precede and/or follow it, and thus its orthographic syllabification). Same character can have a different collation weight based on this context. The character does not change its orthography in these languages. Bangla (and Devanagari), Tamil and Sinhala scripts present a similar scenario, but with vowels. Independent and dependent vowels are sorted differently, the latter normally occurring in syllables with onset consonants. However, in this case the vowels change their orthography as well. Case sensitivity may also be viewed as a form of context sensitivity, where upper case is motivated in the context of proper noun semantics or sentence boundary. The behavior is similar to the change in shape between independent and dependent vowel forms in Bengali, Tamil, Sinahala and other scripts, though this change is motivated under different conditions. This is shown in Cyrillic script as adapted by Mongolian language, as is also true for Latin and Greek scripts. Case also changes the collation of characters in languages using these scripts, including Mongolian. However, collation for all languages is not sensitive to context. Urdu and Sindhi do not have context-dependent variability in collation weights for their characters, even though the context does change the orthography of these characters. This is summarized in the table below.

[2] Though there are some linguistic and lexical differences.

Table 11.1. Influence of Context on Collation and Orthography of Characters in Asian Languages

Language (Script)	Influence of Context on Collation	Influence of Context on Orthography	Remarks
Dzongkha (Tibetan)	Yes	No	Only for Consonants
Lao (Lao)	Yes	No	Only for Consonants
Mongolian (Cyrillic)/ English, French, … (Latin)	Yes	Yes	For both Consonants and Vowels, through Casing
Bengali (Bengali)/ Sinhala (Sinhala)/ Tamil (Tamil)	Yes	Yes	For Vowels
	No	Yes	For Conjunct Consonants for Bengali and Sinhala For Consonant-Vowel Conjuncts for Tamil
Urdu (Arabic)/ Sindhi (Arabic)	No	Yes	For all characters

Asian languages discussed observe multiple levels of collation. From languages like Lao, which have four levels of collation, to languages like Bengali, which mostly collate at a single level, there is complete variety. Urdu and Sindhi have three levels of collation, and Mongolian has two levels of collation. None of the languages show more than four levels of collation.

There are two main challenges faced in modeling the Asian scripts and languages discussed, one orthographic and other technical. Orthographically, many of these languages present a much more complex system [58]. The complexity arises from a variety of factors, including non-monotonic writing system, context-sensitive shaping, variety in orthographic units, with some languages structured around syllables, some on half-syllables, and some on characters.

At the technical level, encoding of these languages in Unicode presents a challenges as it is sometimes arbitrary, and many times redundant, developed through non-academic practical compromises, e.g. to keep backward compatibility, etc. Such decisions have added dimensions in encoding which have to be neutralized through the processes of Normalization, Reordering and Contraction before an encoded string can be sorted. Normalization is usually employed in cases where redundancies have been introduced in the encoding. Reordering is necessary due to monotonic encoding of non-monotonic writing systems (or exceptional ordering in a language, e.g. for marks in French) and contraction is required when composed characters form distinct orthographic entities, different from its parts.

Though these processes are generally defined by Unicode, they still need to be further investigated for each of the languages. This has been addressed to some extent in this volume, but much more work needs to be done in these areas. The recommendations from Unicode provide only a guideline, which work reasonably well for script encoding dealing with a single or a few languages, e.g. Lao, Bengali and Sinhala etc. However, these recommendations have to be reviewed more thoroughly for languages which share the script with many other languages, e.g. Sindhi and Urdu, both using Arabic script, because in these cases the script encoding is a common denominator for all the relevant languages and thus not always effectively catering to the needs of a single one.

Once the input string is processed, in most of the languages, the Unicode collation algorithm is applicable in a straight forward manner. However, one assumption implicit in this algorithm is that sorting is to be performed at word level. This is not true for Dzongkha and Laos, which sort syllable at a time. Thus, the algorithm has to be modified to create collation keys for each syllable and then eventually compared for the syllables in the word. If this is enabled, then the same technique to generate the collation key can be used for each syllable.

In conclusion, Asian scripts present a variety of unique challenges for collation, based on the diversity in scripts in the region. These challenges are caused both by the complexities in the writing systems and by encoding of these systems using Unicode. Complex problems require complex solutions, as in some of the languages discussed. Many of these solutions cannot fit naturally in the existing collation framework defined through the Unicode Collation algorithm, though the current work tries to provide solutions within this framework. However, there may be more natural and simpler solutions if language(s) are encoded differently[3]. This would be a good theoretical exercise, but would be marred by practical issues.

The work presented in this volume is by no means the final word on the collation of these languages. Though the proposed solutions have been tested and the work has been reviewed, more testing is still desired. These algorithms are thus in initial step towards a more rigorous linguistic and computational investigation into these languages.

[3] For example, there is a Chinese standard encoding for Tibetan script which handles collation differently from Unicode.

12. References

[1] Wissink, C. and Kaplan, M. (2003). "Sorting it all out: An Introduction to Collation." *Proceedings of 23rd International Unicode Conference,* Prague, Czech Republic.

[2] Davis, M. and Whistler, K. (2006). "Unicode Collation Algorithm 5.0." Retrieved from http://www.unicode.org/reports/tr10/ on 29th Dec. 2006.

[3] Hussain, S., Durrani, N. and Gul, S. (2006). *PAN Localization Survey of Local Language Computing in Asia 2005.* National University of Computer and Emerging Sciences, Pakistan.

[4] Unicode Consortium (2003). *The Unicode Standard 4.0.* Addison Wesley, New York, USA.

[5] Gillam, R. (2003). *Unicode Demystified.* Addison Wesley, New York, USA.

[6] Bhurgari, A. M. "Enabling Pakistani Languages through Unicode." Retrieved from http://download.microsoft.com/download/1/4/2/142aef9f-1a74-4a24-b1f4-782d48d41a6d/PakLang.pdf on 26th Dec. 2006.

[7] Kenstowicz, M. (1994). *Phonology in Generative Grammar*. Blackwell Publishers, Cambridge, USA.

[8] Afzal, M. and Hussain, S. (2001). "Urdu Computing Standards: Development of UZT 1.01," in the *Proceedings of the IEEE International Multi-Topic Conference*, Lahore, Pakistan.

[9] Hussain, S. and Afzal, M. (2001) "Urdu Computing Standards: UZT 1.01," in the *Proceedings of the IEEE International Multi-Topic Conference,* Lahore, Pakistan.

[10] Wissink, C. (2001). "Issues in Indic Language Collation," in the *Proceedings of 19th International Unicode Conference*, San Jose, USA.

[11] Wikipedia. "Vietnamese Tones." Retrieved from http://en.wikipedia.org/wiki/Vietnamese_Tones on 15th March 2007.

[12] Ethnologue.com. "Bengali: A language of Bangladesh." Retrieved from http://www.ethnologue.com/14/show_language.asp?code=BNG on 22nd March, 2007.

[13] Omniglot.com. "Bengali Alphabet." Retrieved from http://www.omniglot.com/writing/bengali.htm on 22nd March, 2007.

[14] Ishida, R. "Bengali Script Notes". Retrieved from http://people.w3.org/rishida/scripts/bengali/bengali-script/ on 23rd March, 2007.

[15] Bangla Academy (1994). *Bengali-English Dictionary*. Ali, M., Moniruzzaman, M. and Tareque, J. (Eds.). Bangla Academy Press, Dhaka, Bangladesh.

[16] Unicode Consortium. “Default Unicode Collation Element Table” (Allkeys.txt 5.0). Retrieved from http://unicode.org/Public/UCA/latest/allkeys.txt on 23rd March 2007.

[17] Ethnologue.com. “Lao: A language of Laos.” Retrieved from http://www.ethnologue.com/show_language.asp?code=lao on 23rd March, 2007.

[18] Coulmas, F. (1996).*Encyclopedia of Writing Systems*. Blackwell Publishers, Oxford, UK.

[19] Omniglot.com. “Lao Alphabet.” Retrieved from http://www.omniglot.com/writing/lao.htm on 23rd March, 2007.

[20] Phissamy, P., Dalaloy, V., Silimasak, O., Chanhsililath (2007). “Lao Syllabification”, *PAN Localization Working Papers 2004- 2007*. PAN Localization Project, National University of Computer and Emerging Sciences, Lahore, Pakistan.

[21] SEASite Laos, Northern Illinois University. “Lao Vowels.” Retrieved from http://www.seasite.niu.edu/lao/LaoLanguage/LaoAlphabet/LaoVowels.htm on 24th March, 2007.

[22] Bouaravong, P. (2004). “English-Lao & Lao-English Dictionary.”

[23] Aroonmanakun, W. (2002). “Collocation and Thai Word Segmentation.” In *The Proceedings of the Fifth Symposium on Natural Language Processing & The Fifth Oriental COCOSDA Workshop*. Pathumthani: Sirindhorn International Institute of Technology.

[24] Meknawin, S. (1995). “Towards 99.99% Accuracy of Thai Word Segmentation.” Oral Presentation at *The Symposium on Natural Language Processing,* Thailand*.*

[25] Charoenpornsawat, P., Kijsirikul, B. (1998). “Feature-Based Thai Unknown Word Boundary Identification Using Winnow.” In *The Proceedings of the 1998 IEEE Asia-Pacific Conference on Circuits and Systems (APCCAS’98)*.

[26] Unicode Consortium. “Clarification of Bengali Reph and Ya-phalaa.” Retrieved from http://unicode.org/versions/Unicode4.0.1/ on 27th April, 2007.

[27] Davis, M. and Whistler, K. (2006). “Unicode Normalization Forms” Retrieved from http://www.unicode.org/reports/tr15/ on 24th May. 2007.

[28] Ethnologue.com "Dzongkha A Language of Bhutan" Retrieved from http://www.ethnologue.com/show_language.asp?code=dzo on 14th May 14, 2007

[29] Wikipedia "Dzongkha Language" Retrieved from http://en.wikipedia.org/wiki/Dzongkha on 14th May, 2007

[30] Omniglot.com. "Tibetan" Retrieved from http://www.omniglot.com/writing/tibetan.htm on 14th May, 2007

[31] "The Tibetan Language Student" Retrieved from http://www.learntibetan.net/index.htm on 14th May 14, 2007

[32] "The Alphabet" Retrieved from http://www.geocities.com/Athens/Academy/9594/tibet. html on 14th May, 2007

[33] *Geyleg, P* (2007). "Collation in Dzongkha", *PAN Localization Working Papers 2004- 2007*. PAN Localization Project, National University of Computer and Emerging Sciences, Lahore, Pakistan.

[34] *Dzongkha Dictionary*. Dzongkha Development Authority, Thimphu, Bhutan.

[35] "An Ordered List of Collation Elements for Sorting Unicode Dzongkha and Tibetan Data" Retrieved from http://www.dit.gov.bt/guidelines/Dzongkha%20collation%20elements.pdf on 25th May 2007

[36] "Dzongkha Collation Rules" Retrieved from http://www.panl10n.net/Presentations/Cambodia/Pema/Collation(Bhutan).pdf on 25th May 2007

[37] Ethnologue.com. "Mongolian, Halh: A language of Mongolia." Retrieved from http://www.ethnologue.com/show_language.asp?code=khk on 25th March, 2007.

[38] Ethnologue.com. "Mongolian, Peripheral: A language of China." Retrieved from http://www.ethnologue.com/show_language.asp?code=mvf on 25th March, 2007.

[39] Omniglot.com. "Mongolian" Retrieved from http://www.omniglot.com/writing/mongolian.htm on 25th March, 2007.

[40] Altangerel Damdinsuren (2000). *A Modern Mongolian-English Dictionary*. Interpress Publishing, Ulaanbaatar, Mongolia.

[41] Ethnologue.com. “Sindhi: A language of Pakistan.” Retrieved from http://www.ethnologue.com/show_language.asp?code=snd on 26th April, 2007.

[42 Omniglot.com. “Sindhi.” Retrieved from http://www.omniglot.com/writing/sindhi.htm on 26th April, 2007.

[43] Bulchand, D. (1901), revised by Joyo, M. I. (2003). A Manual of Sindhi. Sindhi Language Authority, Hyderabad, Pakistan.

[44] Mewaram, P. (1910). Sindhi-English Dictionary. Reprinted in 1991 by Sindh University Press, Jamshoro, Pakistan.

[45] Ethnologue.com. “Sinhala: A language of Sri Lanka.” Retrieved from http://www.ethnologue.com/show_language.asp?code=sin on 26th April, 2007.

[46] Disanayaka, J. B. (2003). *Say it in Sinhala*. Stamford Lake, Pannipitiya, Sri Lanka.

[47] Omniglot.com. “Sinhala Alphabet” Retrieved from http://www.omniglot.com/writing/sinhala.htm on 25th April, 2007.

[48] Weerasinghe, A. R., Herath, D. L., Gamage, K. (2006). “The Sinhala Collation Sequence and its Representation in Unicode.” In *Localization Focus*, Localization Research Center, University of Limerick, Ireland.

[49] Carter, C. (1924). *Sinhalese-English Dictionary*. Reprinted in 2004 by Asian Educational Services, New Dehli, India.

[50] Samaranayake, V. K., Nandasara, S. T. Disanayaka, J. B., Weerasinghe, A. R. and Wijayawardhane, H. (2003). “An Introduction to Unicode For Sinhala Characters.” *USCS Technical Report 03/01*, Colombo, Sri Lanka.

[51] Ethnologue.com. “Tamil: A language of India.” http://www.ethnologue.com/show_language.asp?code=tam on 26th April, 2007.

[52] Omniglot.com. “Tamil.” Retrieved from http://www.omniglot.com/writing/tamil.htm on 26th April, 2007.

[53] Sura Books (2005). *Tamil-Tamil-English Dictionary*. Sura Books, Anna Nagar, Chennai, India.

[54] LIFCO (2005). *Tamil-Tamil-English Dictionary*. LIFCO, Chennai, India.

[55] Moore, R. (1998). “Introduction to Sinhalese Writing System.” Retrieved from http://www-texdev.ics.mq.edu.au/l2h/indic/Sinhala/lreport/node2.html on 28th April, 2007.

[56] Ethnologue.com. “Urdu.” Retrieved from http://www.ethnologue.com/show_language.asp?code=urd on 26th March, 2007.

[57] Omniglot.com. “Urdu Alphabet.” Retrieved from http://www.omniglot.com/writing/Urdu.htm on 26th March, 2007.

[58] Hussain, S. (2004). “Complexity of Asian Scripts: A Case Study of Nafees Nasta’leeq.” In the *Proceedings of SCALLA*, Kathmandu, Nepal.

[59] Hussain, S. (2004). “Letter to Sound Rules for Urdu Text to Speech System.” In the *Proceedings of Workshop on Computational Approaches to Arabic Script-based Languages,* COLING 2004, Geneva, Switzerland.

[60] National Language Authority. Retrieved from http://www.nla.gov.pk on 7th June, 2007.