Khmer Collation Development

Chea Sok Huor, Atif Gulzar, Ros Pich Hemy and Vann Navy
Csh007@gmail.com, atif.gulzar@gmail.com, pichhemy@gmail.com

Abstract

This document discusses the research on Khmer Standardization (Unicode Collation) project. In order to completely understand the documents, the reader should have some knowledge on Khmer Unicode character characteristics and general collation process. The document is divided into 3 main parts. First, the research methodology that presents the detail of the study including Khmer script ordering analysis and the general algorithm that leads to the Khmer Unicode Collation. Second, the result of the study of the Khmer Unicode Collation project that will provide all the solutions to all the problems faced during the task Finally, the findings and implications of the work that will discuss all the cases that have not been solved during the study.

1. Introduction

The main difficulty for the development of Khmer collation concerns with the source for the collation rule. No exact rule for the order has been found. Until now, only CHUON NATH dictionary is considered official dictionary for Khmer language. Therefore, the dictionary order is adapted. However, CHUON NATH dictionary words are ordered phonetically. In addition, there is no specific rule for the phonetic order in the dictionary. Thus, specific technique needs to be developed for Khmer collation application.

2. Methods

2.1. The complexity of Khmer script ordering

The Khmer word ordering in CHOUNNAT dictionary is very complex because of its phonetic ordering base. Before starting the analysis of the ordering of Khmer script, the first step is to divide the Khmer character into categories and below is the result of the study. Khmer character is categorized into 8 groups:

1. Consonant
2. Independent vowels
3. Dependent vowels
4. Subscript
5. Various sign
6. Digits
7. Khmer digit for divination lore
8. Khmer symbol block

According to CHOUNNAT dictionary, the basic collation sequence of Khmer characters is:

1. Consonant
2. Dependent vowels
3. Subscript

The rest of the categories vary in the sorting sequences and are described as below:

2.1.1. Independent vowel

In CHOUNNAT dictionary, the independent vowels are sorted in the range of the Khmer consonant due to their similar phonetic behavior.

The list given below is in the order of CHOUNNAT dictionary.

1. អុក = អ + ◌ុ + ក
2. ឧកាស=ឧ + ក + ◌ា + ស
3. អុច = អ + ◌ុ + ច

The Khmer word ឧកាស is found among the consonant អ. It is treated as អ + ◌ុ in sorting.

For more information about the ordering of all the independent vowels, see the table of Khmer Independent vowel collation sequences in Table B.1 of Appendix B.

2.1.2. Various sign

In the dictionary, various signs do not affect the order of the words in case there are differences in the base characters. Consider the following order:

1. អឺប = អ + ◌ឺ + ប
2. អ៊ឺយ = អ + ◌ឺ + ◌៊ + យ
3. អឺរ៉ុប = អ + ◌ឺ + រ + ◌៉ + ◌ុ + ប

The various ◌៎ (VISAGNI) is not counted for the comparison process.
(Note: For the order among the various signs, it is considered sorted according to its sound. See the Table B.2 of the Appendix B.)

2.1.3. Digits, Khmer digit for divination lore and Khmer Symbol block

Because the order of Khmer digit, digit for divination lore and Khmer symbol block are not mentioned in the CHUON NATH dictionary and there is no other resource found for its ordering, it is decided to sort them after all the rest of base character categories. The order among the three categories is:

1. Khmer digit
2. Khmer numeric symbol for divination lore
3. Khmer lunar date symbol

2.1.4. Khmer vowels ◌ះ (REAHMUK)

The sorting order of REAHMUK varies in the CHUON NATH dictionary.

- If there is another dependent vowel before the various sign REAHMUK, the REAHMUK is treated as the Khmer consonant ហ (HA) due to its phonetic similarity.
- If the REAHMUK is isolated from other dependent vowel, it is sorted after the vowel ាំ (Khmer vowel SRAK AM)

2.1.5. BA (ប) word lists

As mentioned in the previous section, the various signs are treated as ignorable during the sorting process, but it is an exceptional case for the various sign ◌៉ and ◌៊ (MUUSIKATOAN and TRIISAP) in the ប (BA) word list. It is because of the sound. The order is as follows:

- ប (BA)
- ប៉ = ប + ◌៉
- ប៊ = ប + ◌៊

For example, បា is always smaller than ប៉.

2.2. High Level System Architecture

As illustrated in Figure A.1 of Appendix A, the proposed system is developed by dividing its functionalities into 4 main modules, which will be described later in this section. This division will make the system more reusable and efficient, because each module can be implemented with the technologies best suited to it and the modification of one module does not affect the other, which is easy to control.

2.2.1. External Application

This module focuses on all the user applications that need the functionality of the Collation Engine to compare their Khmer Unicode strings. For example, the sorting engine. The modification of the collation engine does not affect this module.

2.2.2. Collation and Normalization Engine

This part of the system plays a very important role. It is designed to be the common library or assembly that all the external applications will use for their comparison tasks so that they can have a unified collation algorithm for their application. There are two mains modules for this part of the system. First, the Normalization engine is an engine such that its functionality is to normalize the Khmer Unicode word into a common word with respect to Khmer spelling order. Second, the Collation engine is an engine that is responsible for collating the Khmer strings.

2.2.3. Data Reader Engine

This module is an interface between the collation and normalization engine and the data. It prevents those engines from accessing the data directly, which makes thing more complex and unreliable.

2.2.4. Data

This module focuses on all fundamental data that is needed during the process of the collation application. For example, as the collation is based on CHOUN NAT dictionary, its word list must be included in this part.

2.3. General Khmer Collation Algorithm

2.3.1. The requirements and constraints of the collation Engine

The task of the Collation Engine is to give the possibility to compare two Khmer Unicode scripts for the correct order. The requirements of the engine are:

- The collation must give the result of the comparison according to CHUON NATH dictionary.
- For the words that do not exist in CHUON NATH dictionary, the most common rule that is extracted from the dictionary must be applied.

2.3.2. Problem analysis

There are two main problems:

Problem A:
In any circumstance, CHUON NATH dictionary order must be adapted as it is the only official dictionary for Khmer language. However, the words in the dictionary are ordered phonetically. For example:

1. បកតិ = ប + ក + ត + ◌ិ
2. បង់ = ប + ង + ◌់

Visibly, the order must be the first word before the second one as the base consonant ក (KA) is smaller than ង (NGO). Unfortunately, according to CHUON NATH dictionary the order is in contrary. The reason is the pronunciation of the two words.

Solution A:
Since there is no rule for the problem above, the solution is to store all the words of CHUON NATH dictionary so that all the words that exists in the dictionary will be kept as their original order.

Problem 2:

Khmer Unicode font makes it possible for the user to represent a word or syllable in different way.

For example, the word ស្ត្រី can be typed in two different ways:

ស្ត្រី = ស + ◌្ត + ◌្រ + ◌ី
ស្ត្រី = ស + ◌្រ + ◌្ត + ◌ី

In user perspective, the two strings are considered the same according to its visibility. However, in the computer, the two strings are different because of the order.

Solution B
The solution is to do some processing before the comparison. The process is called normalization. The normalized string must respect the spelling order of Khmer language.

2.3.3. General process of Khmer Collation

After the analysis of the problems, the main algorithm for the Khmer Unicode Collation is:

1. Get the two strings to be collated
2. Normalize the two strings into a common form of Khmer word that respect the Khmer spelling order
3. Search for index of the two words in the CHUON NATH dictionary
4. If one of the two normalized words is not found in CHOUNNAT dictionary word lists, apply a rule that is extracted from the dictionary.

Therefore, there are three main processes for Khmer Unicode collation: Khmer Normalization, Get the index of the word from the CHOUNNAT dictionary, and the general term of collation process.

2.3.4. Khmer Unicode Normalization

The main task of the Khmer Unicode Normalization is to standardize those string given into a unique sequence that respect to the Khmer spelling order.

2.3.4.1. General Process of Khmer Normalization

The general processes of the Normalization engine are:

- First, the non-normalized string is converted into a common array of Khmer script and remains the in order of non-normalized string. In this step, the Normalization engine can recognize the nature of the Unicode given whether it is a consonant or a subscript or a vowel etc…
- After getting the array of Khmer script, the Normalization engine will detect each orthographic syllable and arrange them into a normalized string with respect to Khmer spelling order.

2.3.4.2. Khmer Unicode spelling order

The ordering of Khmer spelling is as given below:

1. Base consonant
2. First subscript
3. Second subscript
4. Consonant shifter

5. Vowel
6. Various signs

The Khmer subscript can be divided into 3 categories:

- South subscript: the subscript that is placed below the base character
- West subscript : the subscript that is rendered on the left side of the base character
- East subscript: the subscript that is placed on the right side of the base character

The priorities of the input sequences of the 3 types of subscript are as below:

1. South subscript
2. East subscript
3. West subscript

The independent vowel and digit are considered the stand-alone character, i.e. single orthographic syllable.

2.3.4.3. Syllable detection

The idea to detect syllable is to detect the transition state of each character of the input string. If there is no possible transition for a character, it means the end of orthographic syllable is reached.

The detection of each syllable depends on the possibility of transition between characters. Therefore, it is required to analyze all the possibilities of the input sequence and create possible transition lookup table. For the complete lookup table for the development of Khmer Unicode Normalization and its description, see the Table B.3 of Appendix B.

2.3.5. Khmer Collation element table

After the analysis of some Khmer words in the dictionary, two levels of comparison for Khmer are defined.

1. The first level focuses on all the Khmer base characters. It mainly consists of consonants, subscripts, dependent vowels, independent vowels, Khmer digit, Khmer numeric symbol for divination lore, and Khmer symbol for lunar date.
2. The second level focuses on all the Khmer various signs.

Multiple Mapping:

For Khmer collation element table, it is not always simple as mapping from one character to one collation element. There is case of one to one mapping, one to many mapping (expansion), many to one mapping (contraction) and many to many mapping.

Expansion: Example

ឱ(one single character) and អុ(two characters) are considered to be the same in pronunciation. They are considered the same. However, if they happened to be compared with each other without others, the first one is bigger than the second one.

Consider:

The collation element of អ is [42.0]

The collation element of ◌ុ is [100.0]

So collation element array of អុ must be [42.0], [100.0]

The collation element of ឱ should be [42.0], [100.0], [0.1]

Contraction: Example

All Khmer subscript are represented by 2 Unicode, 17D2 and the Unicode of the base character. However, it has to map into one collation element only.

Many to many mapping: Example

◌ិះ → ◌ិ + ◌ះ

◌ិ + ហ

3. Results

Since the word that exists in the CHUON NATH dictionary are indexed, the performance of Khmer collation for the comparisons those words is 100% accuracy. The problem is to test the extracted rule by comparing the order using the rule and the order in the dictionary.

To test for the performance of the rule, 100 pairs of words in the dictionary were randomly selected. Then, the rule is applied to each pair for its order with each other. After that, the order is compared with the order in the CHUON NATH dictionary.

Within the 100 pairs, 94 pairs result the same as the order in the dictionary. It implies that if CHUON NATH dictionary is necessary for the order, the accuracy of Khmer collation reach 94%.

4. Discussion

4.1. Collation of Khmer date and time

The Khmer collation engine does not provide an option for the collation of Khmer date and time. The reason is not the difficulty of the implementation, but the lack of resource for the format.

4.2. Ordering of book titles

There are some alerts for some specific ordering such as the title of the books or novel or etc. Actually, for that kind of ordering, elimination of some unnecessary words is required during the collation process in order to receive the expected ordering effectively. For example, in English, the words ‘a’ and ‘the’ are eliminated for some searching and sorting purposes in order to receive the most expected result.

As the Khmer collation engine is the base operation for the Khmer standardization, the handling of this case might have a big effect on other usages. Therefore, there should be another collation application that can serve this specific purpose.

4.3. The case of TA and DA in normalization

The case of subscript of TA and DA is still a problem for the implementation of the Khmer Unicode Collation. Actually, the visibility of the two subscripts is the same, but the code is different. Because of its similar visibility, the users usually have some confusion while typing one of the two characters. Therefore, when the users wish to sort their data, they might complain about it as the order must be according to the code of the character.

We attempted to handle the case during the normalization phase by a study on the different usage between the two subscripts, but there is no exact rule for it. Therefore, it was decided to keep it as it is and leave it to the user to be very careful for what they exactly want to type.

5. Conclusion

Since the Khmer word ordering utility is mostly required for everyday use, the project is implemented using two platforms: Microsoft Visual Basic.NET and Java.

The order given is satisfactory. If the user wishes to sort the CHOUNNAT dictionary, it is one hundred percent the same.

Strength: As the project is developed by dividing the complexity in to small modules, each part can be developed using the algorithm best suited. Therefore, the main strength of the application is its reusability. This characteristic is advantageous in case the user expect a new sorting order for their collation. They just modify the collation element table file without the effect overall application.

In addition, the isolating of Khmer Normalization engine from Khmer Collation engine is very advantageous for further usage of the Normalization, as it is not only used for the Collation engine. It will be needed in the Searching utilities, spell checker and so on.

Weakness: The only weakness of the collation application is the speed. If the user wishes to use the collation for sorting a huge amount of data such as database, it will take a long time. The reason is that there are many processes needed in the collation application, first the normalization, second the searching of each string in CHOUNNAT dictionary word list and then if the string is not found the most common rule extracted from CHOUNNAT dictionary is applied.

Future: For the future phase, the need to improve the speed of the collation engine is very necessary, as it is the base functionality for the Khmer standardization. In fact, there is no problem for the sorting of a small amount of data, but if user wishes to sort a huge amount of data such as database, the speed must be more considerable.

In addition, the option to collate Khmer time and date are not yet implemented, because Khmer does not have exact time and date format.

Since the collation engine is implemented as the common assembly or library, the user can use it for their further software development such as sorting of excel sheet or database.

6. References

[1] CHOUNNATH, Dictionnaire Cambodgien, Edition de L'Institut Bouddhique, Phnom Penh, 1967

[2] http://www.unicode.org/charts/PDF/U1780.pdf

[3] http://www.unicode.org/reports/

Appendix A

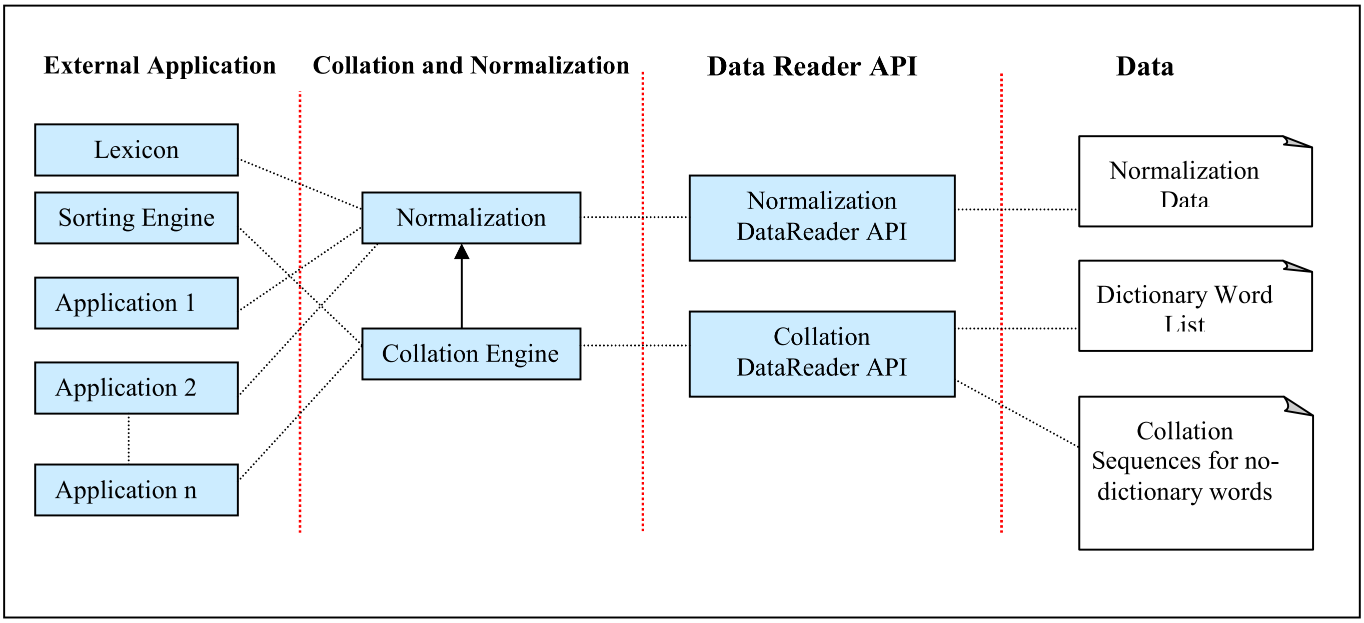

Figure A.1: High Level System Architecture of Khmer Collation

Appendix B

Table B.1: Khmer independent vowel collation sequences

No	Unicode	Character	Substitution	Predecessor
1	17A3	ឣ	ឣ	
2	17A4	ឤ	ឣ + ◌ា	
3	17A5	ឥ	ឣ + ◌ិ	
4	17A6	ឦ	ឣ + ◌ី	
5	17A7	ឧ	ឣ + ◌ុ	
6	17A8	ឨ		ឧ

7	17A9	ឩ	អ +◌ុះ	
8	17AA	ឪ	អ +◌ូ	
9	17AB	ឫ		រ
10	17AC	ឬ		ឫ
11	17AD	ឭ		ល
12	17AE	ឮ		ឭ
13	17AF	ឯ	អ +ែ◌	
14	17B0	ឰ	អ +ៃ◌	
15	17B1	ឱ	អ +ោ◌	
16	17B2	ឲ		ឱ
17	17B3	ឳ	អ +ៅ◌	

Table B.2: Khmer Various Sign Collation Order

No	Unicode	Character
1	17CE	◌៎
2	17D3	◌៓
3	17D9	៙
4	17DA	៚
5	17CA	◌៊
6	17DC	◌ៜ
7	17D8	។ល។
8	17C9	◌៉
9	17C8	◌ៈ
10	17CC	◌៌
11	17D4	។

12	17D5	៕
13	17CB	◌់
14	17D7	ៗ
15	17CF	◌៏
16	17D1	◌៑
17	17D0	◌័
18	17D6	៖
19	17D2	◌្
20	17D8	៘
21	17DC	ៜ
22	17DD	៝

Table B.3 : Transition table for detecting Khmer syllable

	C	CS	WV	NV	SV	EV	WSS	ESS	SSS	VS
C	0	1	1	1	1	1	1	1	1	1
CS	0	0	1	1	1	1	1	1	1	1
WS	0	0	0	0	0	0	1	1	1	1
NV	0	0	0	0	0	0	1	1	1	1
SV	0	0	0	0	0	0	1	0	0	1
EV	0	0	0	0	0	0	1	0	0	1
WSS	0	1	1	1	1	1	0	1	1	1
ESS	0	1	1	1	1	1	1	0	1	1
SSS	0	1	1	1	1	1	1	0	0	1
VS	0	0	0	0	0	1	1	0	1	1

Shortcut	Abbreviation
C	Consonant
CS	Consonant Shifter
WV	West Vowel
NV	North Vowel
SV	South Vowel
EV	East Vowel
WSS	West Subscript
SSS	South Subscript
ESS	East Subscript
VS	Various Sign

Initial states are: **C**

<u>**Note:**</u> In the transition table:

- 0 means ***NOT Possible transition*** and 1 means ***possible transition.***
- The current state is in ROW.
- The transition state is in COLUMN.