

General Structure Based Collation of Tibetan Syllables

Heming HUANG^{1,2,†}, Feipeng DA¹

¹ School of Automation, Southeast University, Nanjing, Jiangsu 210096, China;

² Physics Department, Qinghai Normal University, Xining, Qinghai 810008, China

Abstract

Tibetan syllable collation is the foundation of Tibetan information retrieval, machine translation, spelling checking, and corpus creation etc. This paper proposes a Tibetan syllable collation algorithm based on the idea of expanding two-dimensional syllable into one-dimensional letter series by introducing a general syllable structure. This general syllable structure consists of base consonant (BC), head consonant (HC), prefix consonant (PC), the first foot consonant (FC1), the second foot consonant (FC2), the first vowel (V1), the second vowel (V2)/the first suffix consonant (SC1), and the second suffix consonant (SC2). Before comparing with others, a syllable should be expanded to the series of BC, HC, PC, FC1, FC2, V1, V2/SC1, and SC2. If there is no letter in a particular position of the syllable, the corresponding position of the expanded series should be filled with a special character whose collation element is less than that of any Tibetan letters. Furthermore, while expanding, if a syllable has both prefix consonant and head consonant, the corresponding letters ཨ, ལ, and ས in the expanded series should be substituted by ཨ, ལ, and ས respectively. Thus, each syllable could be expanded to a canonically equivalent letter series and comparing of two syllables becomes comparing of two letter series.

Key words: Tietan Syllable; Collation; General Structure; Transliterating letter.

1. Introduction

Collation is a general term for the process and function of determining the sorting order of character strings. It is a key function in computer systems; whenever a list of string is presented to users, they are likely to want it in a sorted order so that they can easily and reliably find individual strings. Thus, it is widely used in human-machine interfaces. It is also crucial for the operation of databases, not only in sorting records but also in selecting sets of records with fields within given bounds [1]. Tibetan syllable collation, generally speaking, is the foundation of Tibetan information retrieval, machine translation, spelling checking, and corpus creation etc.

Research on Tibetan syllable collation is lagging behind other languages. It may be attributed to the later start, the special characteristic of Tibetan syllables, and the lack of support in terms of manpower, funding etc. There are few papers in this field, and Tibetan syllable collation is far from being completely resolved.

This paper proposes a Tibetan syllable collation algorithm by introducing a general syllable structure and, based on this general structure, expanding each two-dimensional syllable into one-dimensional letter series. The rest of paper is organized as follows. Section 1 introduces Tibetan script briefly while section 2 summarizes the previous work in this field. Section 3 abstracts a general Tibetan syllable structure and

[†] Corresponding author.

Email addresses: hbm-2008@qq.com (Heming HUANG), fpda@seu.edu.cn (Feipeng DA).

section 4 describes preprocessing for some exceptive syllables so that all Tibetan syllables and Tibetan transliterating composed characters could be represented by the general structure directly or indirectly. After that, section 5 introduces the method of expanding two-dimensional syllable into one-dimensional letter series. Finally, section 6 gives conclusions and tracks for future work.

2. Brief Introduction to Tibetan Script

The Tibetan script, which has a history of 1,500 years, is used for writing Tibetan in several countries and regions throughout the Himalayas. Aside from Tibetan area of People's Republic of China, the script is used in Nepal and northern areas of India bordering Tibet where large Tibetan-speaking populations reside. The Tibetan script is also used in Bhutan to write Dzongkha, the official language of that country. In addition, Tibetan is used as the language of philosophy and liturgy by Buddhist traditions spread from Tibet into the Mongolian cultural area [2].

2.1. Native Tibetan Syllable

There are 30 consonants and 5 vowels in native Tibetan script. The 30 consonants are: ཀའ་ཀའ་ ཅའ་ཅའ་ ཉའ་ཉའ་ ཏའ་ཏའ་ ཐའ་ཐའ་ ཌའ་ཌའ་ ཎའ་ཎའ་ སའ་སའ་ སྐའ་སྐའ་ སྒའ་སྒའ་ སྔའ་སྔའ་ སྨའ་སྨའ་ སྩའ་སྩའ་ སྨྱའ་སྨྱའ་ སྨྲའ་སྨྲའ་ སྨླའ་སྨླའ་ སྨྤའ་སྨྤའ་ སྨྦའ་སྨྦའ་ སྨྦྷའ་སྨྦྷའ་ སྨྨའ་སྨྨའ་ སྨྩའ་སྨྩའ་ སྨྪའ་སྨྪའ་ སྨྫའ་སྨྫའ་ སྨྫྷའ་སྨྫྷའ་ སྨྭའ་སྨྭའ་ སྨྯའ་སྨྯའ་ སྨྱའ་སྨྱའ་ སྨྲའ་སྨྲའ་ སྨླའ་སྨླའ་ སྨྤའ་སྨྤའ་ སྨྦའ་སྨྦའ་ སྨྦྷའ་སྨྦྷའ་ སྨྨའ་སྨྨའ་ སྨྩའ་སྨྩའ་ སྨྪའ་སྨྪའ་ སྨྫའ་སྨྫའ་ སྨྫྷའ་སྨྫྷའ་ སྨྭའ་སྨྭའ་ སྨྯའ་སྨྯའ་. Tibetan letters can combine with others both horizontally and vertically. The combination of letters, allowed by gender rule, is called syllable. There is a base consonant (BC) in each syllable. Other consonants, according to their relative position to the base consonant, are called prefix consonant (PC), head consonant (HC), foot consonant (FC), the first suffix consonant (SC1) and the second suffix consonant (SC2) respectively (Fig. 1). Base consonant, head consonant (if any), foot consonant (if any), and vowel is stacked vertically, and the stack is called Tibetan composed character. When a vowel combines with a consonant stack, only the vowel mark is written. The marks of vowels ཨ, ཉ, མ, and ས are ཨ, ཉ, མ, and ས respectively. Three vowel marks ཨ, ཉ, and མ are put above the consonant stack while one of them ཨ is put below. The absence of vowel mark implies that the first vowel sound (like a short “ah” in English) is present and is not modified to one of the four other possibilities [2].

Tibetan characters have no such variation as uppercase and lowercase. However, some consonants, when they are put above/below base consonant, have variant shape(s). For example, when the letters ཡ, ར, and ལ are put below base consonant, their variations are ཡ, ར, and ལ respectively. And when the letter ར is put above base consonant, its variation is ར.



Fig.1 A Typical Tibetan Syllable

Native Tibetan uses inter-syllable mark " " for separating syllables and uses the symbol " " for marking the end of a phrase or sentence. For example, there are five inter-syllable marks in native Tibetan phrase

2.2. Tibetan Transliterating Composed Character

To represent Indic words correctly, some special letters are introduced to Tibetan script. These special letters, together with some of native Tibetan letters constitute Tibetan Transliterating Alphabet. There are 34 consonants and 16 vowels in this new Alphabet (Fig. 2.b). The 34 consonants are: ཀལག་ཁྲང་། ཅཿཇཱླུ། རཐ་རྟམ། ཉཻད་ནྭམ། པ་ཕ་བ་ཏྲམ། ཡ་ར་ལ་ཤ། སཱེཔ་མ་ད། གྲ། and the 16 vowels are: ཨ་ཞ། ཨི་ཞ། ཨུ་ཞ། འི་ཞ། འི་ཞ། ཨེ་ཞ། ཨོ་ཞ། ཨོ་ཞ།. In the same way, only the vowel mark is written when a vowel is put to a consonant stack. The corresponding marks of all transliterating vowels ཨ་ཞ། ཨི་ཞ། ཨུ་ཞ། འི་ཞ། འི་ཞ། ཨེ་ཞ། ཨོ་ཞ། are ཁ་ལྷ། འི་ཞ། འི་ཞ། འི་ཞ། འི་ཞ། འི་ཞ། འི་ཞ། respectively. The symbol "□" means, when the vowel ཨ་ཞ combines with consonant, there is no corresponding mark.

a. Native Tibetan Letters b. Tibetan Transliterating Letters

According to Tibetan orthography, a native syllable could have at most three consonants stacked vertically. The orthography also gives some strict rules on which letters can appear on a particular position of a syllable. For example, only letters འ, ཡ, and ས could appear on head position and only letters ལ, ར, ལ, and ས could appear on foot position etc. Nevertheless, it is freer for the stacking of Tibetan transliterating letters. Stacks of four or five consonants, as shown in Fig.3, could be found in Chinese national standard of Tibetan Coded Character Set-Extension A and B every once in a while [3][4]. The stack of Tibetan transliterating letters is called Tibetan transliterating composed character and abbreviated as TTCC in this paper.

the consonants are: ��ྒལ་ལྷ་བླ་བ། ཅམ་ཅ་ཉ། ར་ཐ་ཏ་ཏྟ། ཏ་ཐ་ད་ན་ན། ལ་ལ་ལ་ན་ན། ཅ་ཚ་ཅ་ཚ་ལ། ར་ལ་ལ་ལ། ལ་ལ་ལ་ལ། ཏ་ལ། ལ། and the vowels are: ཨ་ཨ། ཨ་ཨ། ཨ་ཨ། ཨ་ཨ། ཨ་ཨ། ཨ་ཨ། ཨ་ཨ།.



Fig.3 A Tibetan Transliterating Composed Character of Five Consonants

3. Literature Survey

As mentioned above, Tibetan syllable is a two-dimensional composition of Tibetan letters. To compare two syllables, each of them should be expanded to a one-dimensional letter series. However, two problems should be resolved before expanding. One is whether there is a general structure for all Tibetan and the other is what the reasonable expanding order is if there is such a general syllable structure.

In refs. [7-9], Jiang Di et al. have proposed that all Tibetan syllables can be classified into four types with each of them has a common structure. As shown in Fig.4, the first type are syllables that have neither head consonant nor prefix consonant; the second and third types are syllables that have either head consonant or prefix consonant while the fourth type are syllables that have both head consonant and prefix consonant. Comparing of two syllables depends not only on the letters in a syllable but also on the syllable's structure. Jiang Di et al. have proposed that, at first, each syllable's base consonant should be compared with each other. If they are different, two syllables' order could be decided by their base consonants directly. Otherwise, two syllables' order should be decided by the syllable structures and the letters in the syllable. If their structures are different, the two syllables' order could be decided by the structures' order. The order of these four structures is 4.a, 4.b, 4.c and 4.d. For example, syllables ལྷ་བླ་བ།, ལྷ་བླ་བ།, ལྷ་བླ་བ།, and ལྷ་བླ་བ། have the same base consonant and different structures, their order could be determined by the structures. If both base consonants and structures are the same, their order could be decided by the letters on the next position of each syllable. This next position is different in different structures. It is foot consonant in structure (a), and it is prefix consonant, head consonant, and prefix consonant in structure (b), (c), and (d) respectively.

Generally speaking, a syllable's collation is determined orderly by the sequence BC, PC, HC, FC, V, SC1, and SC2 as shown in Fig.5. However, different structure corresponds to different sequence because of the absence of head consonant or/and prefix consonant. For example, to structure 4.a, this sequence would be BC, FC, V, SC1, and SC2 while to structure 4.b, this sequence would be BC, PC, FC, V, SC1, and SC2.

It is obvious that the proposition of four structures in Refs. [7]-[9] seems less methodical and makes Tibetan collation algorithm perplexing.

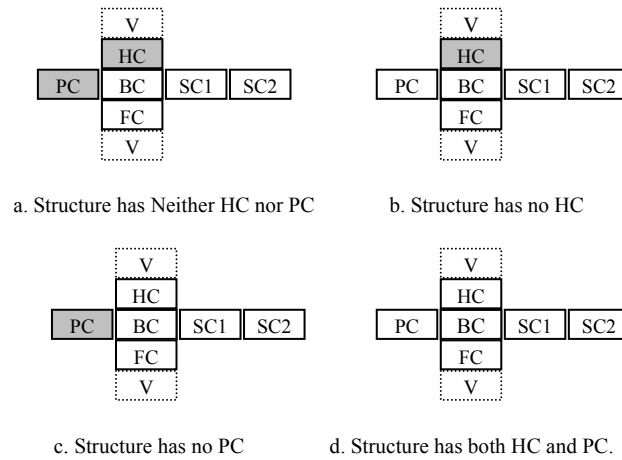


Fig. 4 Tibetan Syllables' Structures.

In Ref. [10], Huang et al. have proposed that all Tibetan syllables should be represented by a common structure as depicted in Fig. 4(d). Moreover, every syllable should be expanded to a letter string as the sequence prescribed by Fig.5. If a syllable has no letter on a particular position, the corresponding vacant position in the expanded series should be filled with a special character whose collation element is less than that of any Tibetan letters. In this paper, the symbol "□" is temporally used to fulfill the vacant position. For example, syllables གྲུབས and བཟུབས should be expanded to letter strings གྲུབས and བཟུབས respectively. By comparing letter strings གྲུབས and བཟུབས, two syllables གྲུབས and བཟུབས could be compared. In fact, except syllables as depicted in Fig. 4(c), all Tibetan syllables could be compared in this way. When the syllables with the structure Fig. 4(c) compare with the syllables with the structure Fig. 4(b) and Fig. 4(d), the vacant prefix position of the Fig. 4(c) should be padded with letter ཨ and ཁ respectively.

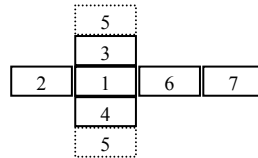


Fig. 5 A Typical Syllable's Collation is Determined Orderly by the Sequence BC, PC, HC, FC, V, SC1, and SC2.

Comparing with perplexing four-structures-proposition of refs [7]-[9], the one-structure-proposition of Ref. [10] is more perspicuous but there are still many exceptive syllables. So a more comprehensive general structure is expected.

4. General Tibetan Syllable Structure

According to the related statements of technical reports of Unicode and ISO [11] [12], in order to resolve Tibetan character collation completely, eight kinds of string's collation should be discussed. These eight kinds of string are common native syllable, grammar affected native syllable, TTCC, non-syllable Tibetan letter string, foreign letter string, Tibetan symbol string, Tibetan digit string, and the mixed string of above. Among them, the collation of common native Tibetan syllable, grammar affected syllable, and TTCC have the precedence over that of others, since Tibetan dictionaries have obvious and strict rules on them. Based on the one-structure-proposition of Ref. [10], the following sections will abstract a more comprehensive

4.1. Double Vowel in Some TTCCs

It is obvious that transliterating syllables have two vowels. For example, syllable 𐌂𐌰 have two vowels 𐌂 and 𐌰, and syllable 𐌂𐌴 have two vowels 𐌂 and 𐌴. Therefore, it is necessary to differentiate a transliterating syllable's vowels as the first vowel (V1) and the second vowel (V2). Among 16 transliterating vowels, the first 14 vowels 𐌂, 𐌄, 𐌆, 𐌈, 𐌊, 𐌋, 𐌍, 𐌏, 𐌑, 𐌒, 𐌔, 𐌖, 𐌘, 𐌚, and 𐌜 could only be the first vowel while the vowels 𐌂, 𐌄, and 𐌴

could only be the second vowel. Some transliterating syllables just have one vowel mark, which means the other vowel is \mathfrak{a} . For example, to syllable \mathfrak{a} , the first vowel is \mathfrak{a} and the second vowel is \mathfrak{a} . Similarly, to syllable \mathfrak{a} the first vowel is \mathfrak{a} and the second vowel is \mathfrak{a} . It is obvious from the rule B.1 that if two transliterating syllables have the same consonants, their order is decided by the first vowel; if their first vowels also are the same, their order is decided by the second vowel.

4.2. Double Foot Consonant in Some TTCCs

A few native Tibetan syllables have double foot consonant. For example, syllable \mathfrak{a} has double foot consonant \mathfrak{a} and syllable \mathfrak{a} has double foot consonant \mathfrak{a} . In fact, in Tibetan Character Set-Extension A and B, which have collected roughly 6400 TTCCs, many TTCCs have double foot consonant. These double foot consonants, their decomposition, and related examples are listed in Table 1.

Table 1 Double Foot Consonant in Some TTCCs

Dual Foot Consonant	1 st Foot Consonant	2 nd Foot Consonant	Examples
\mathfrak{a}	\mathfrak{a}	\mathfrak{a}	\mathfrak{a} \mathfrak{a} \mathfrak{a}
\mathfrak{a}	\mathfrak{a}	\mathfrak{a}	\mathfrak{a} \mathfrak{a} \mathfrak{a}
\mathfrak{a}	\mathfrak{a}	\mathfrak{a}	\mathfrak{a} \mathfrak{a} \mathfrak{a}
\mathfrak{a}	\mathfrak{a}	\mathfrak{a}	\mathfrak{a} \mathfrak{a} \mathfrak{a}
\mathfrak{a}	\mathfrak{a}	\mathfrak{a}	\mathfrak{a} \mathfrak{a} \mathfrak{a}
\mathfrak{a}	\mathfrak{a}	\mathfrak{a}	\mathfrak{a} \mathfrak{a} \mathfrak{a}

DUCET (Default Unicode Collation Element Table) gives collation element to every foot consonant, such as \mathfrak{a} , \mathfrak{a} , and \mathfrak{a} , not to their combination such as \mathfrak{a} and \mathfrak{a} [1]. Therefore, it is better to differentiate double foot consonant as the first foot consonant (FC1) and the second foot consonant (FC2).

4.3. General Tibetan Syllable Structure

As shown in Fig. 1, a typical native Tibetan syllable has seven letters: BC, HC, PC, FC, V, SC1, and SC2. In addition, as discussed in section 3.1 and section 3.2, some TTCCs have two vowels V1 and V2 and some TTCCs have two foot consonants FC1 and FC2. Therefore, the general syllable structure should include nine elements BC, PC, HC, FC1, FC2, V1, V2, SC1, and SC2. Nevertheless, the second vowel and the first suffix consonant could share a common position since one of the second vowels \mathfrak{a} appears in the very position of SC1 and there is no second vowel in native Tibetan syllable. Thus, the general syllable structure should have eight elements: BC, HC, PC, FC1, FC2, V1, V2/SC1, and SC2 as shown in Fig.6.

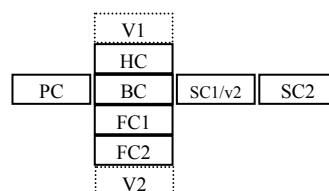


Fig.6 Tibetan Syllable's General Structure

5. Preprocessing for Some Exceptions

The general structure described by Fig.5 could represent directly the common native Tibetan syllables and TTCCs tallying with the definition of transliterating syllable. However, some grammar-affected native syllables and TTCCs could not be represented by this general structure directly. Therefore, a reasonable preprocessing is needed to these syllables and TTCCs.

5.1. Preprocessing for Grammar-affected Syllables

Tibetan grammar affects Tibetan syllable's spelling and this affection usually happens at the suffix positions. Under the affection of grammar, some syllables even have their structures changed. Therefore, grammar-affected syllables can be divided into two classes: syllables that could be represented by the general structure directly and those that could not.

5.1.1. Grammar-affected Syllables that Could be Represented by the General Structure Directly

Grammar-affected syllables that could be represented by the general structure directly could be further divided into two classes. The first class is syllables that the letters on the suffix positions tally with Tibetan orthography. For example, after adding the auxiliary word ྤ that denotes casual relationship, to syllable མཐུག, we have syllable མཐུགྤ and the letter ྤ in this syllable tallies with Tibetan orthography's restriction to the first suffix consonant. The second class is syllables in which the letters on suffix position do not tally with Tibetan orthography's restriction. However, they could be considered suffix consonants intentionally. For example, letters འ and ྤ in the auxiliary word འ of syllables རྩའ and རྩའྤ could be treated as two suffix consonants. Similarly, letters འ and མ in auxiliary word འམ of syllables ལའམ, རྩའམ, and རྩའམྤ could be treated as suffix consonants. The above two kinds syllables although affected by the grammar but have not changed their structure, therefore, they could be represented by the general structure directly.

5.1.2. Grammar-affected Syllables that Could be Represented by the General Structure after Reasonable Segmentation

In Tibetan, some grammar-affected syllables especially those affected by the auxiliary words have changed their structure greatly. These syllables could be classified to four kinds. The first kind is syllables affected by the auxiliary word འེ such as མཐུགའེ and རྩའེ. The second kind is syllables affected by the auxiliary word འུ such as རྩའུ and རྩའུྤ. The third kind is syllables affected by the auxiliary word འུྤ such as རྩའུྤ, རྩའུྤྤ, རྩའུྤྤྤ, and རྩའུྤྤྤྤ. The last kind is syllables affected by the auxiliary word འུྤ and འེ simultaneously such as རྩའུྤའེ.

Different dictionaries treat the auxiliary words འེ, འུ, and འུྤ differently. Some dictionaries consider each auxiliary word an independent syllable while others consider consonant འ as the first suffix consonant and vowels འེ, འུ, or འུྤ as the second suffix consonant. For the sake of simplicity, this paper treats each auxiliary word འེ, འུ, and འུྤ as an independent syllable. Thus, in collation, such syllable as མཐུགའེ should be segmented to syllable series མཐུགའེ, and such syllable as རྩའུྤའེ should be segmented to syllable series རྩའུྤའེ. It is obvious that, after this reasonable segmentation, every syllable in the series could be represented by the general structure. In another word, all such syllables as མཐུགའེ, རྩའུྤའེ etc. could be represented by the general structure indirectly.

5.2. Preprocessing for Some TTCCs

As mentioned above, a transliterating syllable at most is composition of base consonant, the first foot consonant, the second foot consonant, the first vowel, and the second vowel. Some of TTCCs themselves are syllables. For example, character རྩ, which consists of base consonant ར, the first vowel རྩ, and the second vowel རྩ, is a syllable obviously. Other TTCCs that do not tally with the definition of syllable could be segmented to syllable series. The procedure is as follows:

If the current TTCC is not a syllable, extract and consider the first layer consonant as an independent syllable. Then judge whether the remnant part is a syllable. If it is, we have a syllable series; otherwise, continue this process until the remnant part tallies with the definition of syllable. Thus, any TTCCs could be decomposed to a syllable series. For example, extract the first layer consonant ར of the TTCC རྩ and consider it as an independent syllable, the remnant part རྩ is a syllable. Thus, the TTCC རྩ is decomposed to syllable series རྩ. Similarly, the TTCC shown in Fig.3 could be decomposed to syllable series རྩ. Thus, all TTCCs could be represented by the general syllable structure directly or indirectly.

6. Expanding Order of the General Syllable Structure

As discussed above, all native Tibetan syllables and TTCCs could be represented by the general structure directly or indirectly. Thus, the foundation is the collation of syllables that could be represented by the general structure. To collate such a syllable, it should be expanded to letter series at first. As depicted in Fig.7, each syllable is expanded to the sequence of BC, HC, PC, FC1, FC2, V1, SC1/V2, and SC2. In Tibetan syllable, the vacant positions also involve sorting. So if there is no letter in a particular position, the corresponding position in the expanded series should be padded with a special character whose collation element is less than that of any Tibetan letters.

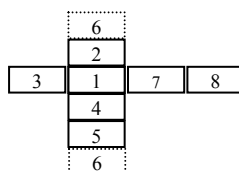


Fig.7 The Expanding Order of General Structure

Thus, each syllable could be expanded to a letter series and the syllables' order could be decided by this letter series. Table 2 shows the expanded series of 20 typical syllables with the same base consonant ར. The odd column of the table is the syllables sorted in their dictionary order while the even column is their corresponding letter series.

By careful studying of Table 2, it could be found that, except syllables རྩ, རྩ, and རྩ, all other syllables' collation could be decided by their expanded series directly. The common feature of syllables རྩ, རྩ, and རྩ is that they have both head consonant and prefix consonant. According to Tibetan orthography, if a syllable has the prefix and head consonant simultaneously, the prefix consonant must be ར and the head consonant must be ར, ར, or ར. In column 5 of table 2, the six syllables' order would be རྩ, རྩ, རྩ, རྩ, རྩ, and རྩ if they were decided simply by their expanded form. However, their dictionary order is རྩ, རྩ, རྩ, རྩ, རྩ, and རྩ. That is to say, the syllables that have both prefix consonant and head consonant should be collated after other syllables. In another words, the comparison of these syllables is context sensitive: it depends on more than just single characters compared directly against one another.

[illegible]

7. Conclusion and Future Works

Based on the previous study, this paper proposes a general Tibetan syllable structure that consists of BC, HC, PC, FC1, FC2, V1, V2/SC1, and SC2 as shown in Fig.5. It could represent all native Tibetan syllables and TTCCs directly or indirectly. A syllable, before comparing with other syllable, should be expanded to a series in the sequence of BC, HC, PC, FC1, FC2, V1, V2/SC1, and SC2 as shown in Fig.6. If there is no letter in a particular position, the corresponding position in the expanded series should be filled with a special character whose collation element is less than that of any Tibetan letters. Furthermore, if a syllable has both prefix consonant and head consonant, the corresponding letters ར, ལ, and ས in the expanded series should be substituted by ར, ལ, and ས respectively. Thus, comparing of two syllables or transliterating characters becomes comparing of two letter series.

Our future work will focus on the following two aspects. The first is revision of some Tibetan characters' collation elements. To a few transliterating letters, their collation elements are not so reasonable and a careful revision is needed. The second is the compression of collation element. The collation element of a letter consists of four fields and each field is a 2-byte hexadecimal integer. For example, the collation element of the first Tibetan letter ཀ is [.1C22.0020.0002.0F40]. To a syllable, If collation element is assigned to every letter in the expanded series, it need 56 bytes. So the compression is very necessary.

Acknowledgement

This work is supported by the National Natural Science Foundation of China under Grant No.60973040; the Yong Scholar's Research Foundation of Qinghai Normal University.

References

- [1] Mark Davis, Ken Whistler. Unicode collation algorithm[M]. <http://www.unicode.org/reports/tr10/tr10-20.html>.
- [2] The Unicode Consortium. Unicode Standard[M]. Boston: Addison-Wesley, August 2003, pp.343-353.
- [3] State Quality Supervision Bureau of China. Information technology-Tibetan coded character set-extension A[S]. Beijing: China Standards Publishing House, 2007-5.
- [4] State Quality Supervision Bureau of China. Information technology-Tibetan coded character set-extension B[S]. Beijing: China Standards Publishing House, 2009-1.
- [5] Zhang Yisun. The Tibetan-Chinese Dictionary[M]. Beijing: The Ethnic publishing house, 1985-7.
- [6] Sarat Chandra Das. A Tibetan-English Dictionary with Sanskrit Synonyms[M]. New Delhi: South Asia Books, 2005.
- [7] Jiang Di, Zhou Jiwen. On the sequence of the Tibetan words and the methods of making sequence[J]. Journal of Chinese Information Processing, 14(1) (2000) 56-64.
- [8] Jiang Di, Kang Caijun. The sorting mathematical model and algorithm of written Tibetan language[J], Chinese Journal of computer, 20(4) (2004) 524-529.
- [9] Lin Heshui, Cheng G Wei, Cao Hui, et al. A method for ordering Tibetan text in accord with ISO 14651[J]. Journal of Chinese Information Processing, 18(5)(2004) 36-41.
- [10] Huang Heming, Zhao Chenxing. A DUCET-based Tibetan Sorting algorithm[J]. Journal of Chinese Information Processing, 22(4) (2008) 56-64.
- [11] ISO/IEC, "Method for comparing character string and description of the common Template Tailorable ordering", <http://www.iso.org/iso/>
- [12] ISO/IEC. Specification method for culture conventions[S]. www.cl.cam.ac.uk/~mgk25/volatile/iso-14652.pdf.
- [13] An Shixing. The Sanskrit-Tibetan-Chinese Dictionary[M]. Beijing: The Ethnic publishing house, 1991-4.
- [14] Tibetan compilation and translation bureau of Aba Prefecture. The Sanskrit-Tibetan Dictionary. Lanzhou: Gangsu Ethnic publishing house, 1996.