Collation of Transliterating Tibetan Characters

Heming Huang[1,2,*] and Feipeng Da[1]

[1] School of Automation, Southeast University, Nanjing, Jiangsu 210096, China
[2] School of computer Science, Qinghai Normal University, Xining, Qinghai 810008, China
huang-heming@sohu.com, dafp@seu.edu.cn

Abstract. The transliterating Tibetan characters used specially to transliterate foreign scripts have two collations: collation with the rules of native Tibetan dictionary and with that of transliterating Tibetan dictionary. This paper proposes two general structures for transliterating characters. Based on these general structures, a collation scheme is developed so that all transliterating characters can be collated correctly and effectively.

Keywords: Tibetan, character, collation, structure.

1 Introduction

The Tibetan script is an alphasyllabary, a segmental writing system in which consonant-vowel sequences are written as a unit. Tibetan has two alphabets: the native Tibetan alphabet used in daily life of Tibetan people and the transliterating Tibetan alphabet used specially to transliterate foreigner scripts especially the Sanskrit.

The native Tibetan has 30 consonants and four vowels. The 30 consonants are 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, and ཨ while the four vowels are ི, ུ, ེ, and ོ. In addition to native Tibetan alphabet, a special alphabet was invented so as to transliterate the Buddhist scripture written in Sanskrit. Because of the exact one-to-one correspondence between this special invented alphabet and the Sanskrit alphabet, any Sanskrit sentence can be converted into the Tibetan sentence exactly and vice versa. This transliterating alphabet has 34 consonants and 16 vowels. The 34 consonants are ཀ, ཁ, ག, གྷ, ང, ཙ, ཚ, ཛ, ཛྷ, ཉ, ཊ, ཋ, ཌ, ཌྷ, ཎ, ཏ, ཐ, ད, དྷ, ན, པ, ཕ, བ, བྷ, མ, ཡ, ར, ལ, ཝ, ཤ, ཥ, ས, ཧ, and ཀྵ, and the 16 vowels are ཨ, ཨཱ, ཨི, ཨཱི, ཨུ, ཨཱུ, ཨྲྀ, ཨཷ, ཨླྀ, ཨཹ, ཨེ, ཨཻ, ཨོ, ཨཽ, ཨཾ, and ཨཿ. As shown in Fig. 1, the two alphabets are different but they share 28 letters ཀ, ཁ, ག, ང, ཅ, ཇ, ཉ, ཏ, ཚ, ཛ, ཝ, ཞ, ཟ, འ, ཡ, ར, ལ, ཤ, ས, ཧ, ཨ, ི, ུ, ེ, and ོ in common.

The transliterating Tibetan is different from the native Tibetan in many ways. One difference is that the transliterating Tibetan has two kinds of collation. The first kind is that all the characters need to be collated are just the transliterating characters and

* Corresponding author.

M. Zhou et al. (Eds.): NLPCC 2012, CCIS 333, pp. 78–84, 2012.

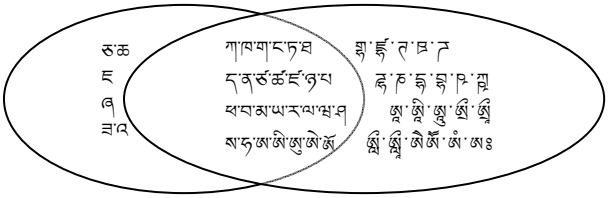

Fig. 1. The native Tibetan alphabet (left) and the transliterating Tibetan alphabet (right)

the collation of these characters follows the sorting rules of the transliterating Tibetan dictionary [1]. The second kind involves the collation of both the transliterating characters and the native Tibetan syllables, and the collation follows the sorting rules of the native Tibetan dictionary [2].

Di Jiang et al. have proposed a sorting algorithm of Tibetan script [3]. Heming Huang et al. have evaluated each Tibetan letter or symbol with the Unicode collation element and proposed a general structure for collation of Tibetan syllables [4-5]. As a matter of fact, they just collate both the transliterating characters and the native Tibetan syllables with the sorting rules of the native Tibetan dictionary. Although, not used so frequently, the transliterating characters are great in number: there are more than 6600 transliterating Tibetan characters [6-7]. By far, it is still an open problem to collate the transliterating Tibetan characters with the sort rules of the transliterating Tibetan dictionary.

2 The Judgment of the Transliterating Tibetan Characters

To realize the second kind collation accurately, it is necessary to distinguish the transliterating characters from the native Tibetan syllables correctly.

It is easy to distinguish a transliterated sentence from a native Tibetan sentence. As shown in Fig. 2, there is an inter-syllable separator '·' between every two syllables of a native Tibetan sentence while there is no such separator in a transliterated sentence.

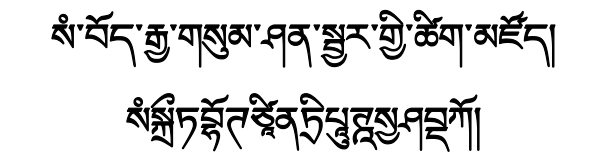

Fig. 2. A native Tibetan sentence (row 1) versus a transliterated sentence (row 2)

Some transliterated characters or phrases are used as a common syllable in a native Tibetan sentence. For example, the first syllable "ཨཾ" in the first row of Fig. 2 is a transliterating character; however, it is separated by '·' from others as if it is a native syllable. Therefore, it is not easy to distinguish a transliterating character from the

native Tibetan syllables under this circumstance. It should be judged with the native Tibetan orthography. Generally, a pre-composed character is a transliterating character if it meets one of the following conditions.

1) A pre-composed character has the transliterating vowel ི, ི, ུ, ུ, ྲྀ, ཷ, ླྀ, ཹ, ེ, ོ, ཻ, or ཽ.

2) A pre-composed character has the diacritic sigh ྀ, ཱ, ྃ, ྄, ྅, ཾ, or ཿ.

3) A pre-composed character has the transliterating consonants ཀ, ཁ, ཏ, ཐ, ཅ, ཆ, ཇ, ཉ, ཀྵ, ཀྐ, or ཀྱ.

4) A pre-composed character has two consonants, but the first consonant is none of ར, ལ, and ས while the second consonant is none of ྱ, ྲ, ླ, and ྭ. Examples of such characters are ཀྱེ, ཀྲ, ཀླ, and དྲི.

5) A pre-composed character has three consonants, but the first one is none of ར, ལ, and ས while the third one is none of ྱ, ྲ, ླ, and ྭ. Examples of such characters are ཀྲུ, ཀྱུ, སྐྱེ, and ཀྲི.

6) A pre-composed character has more than three consonants. Examples of such characters are སྐྲུ, ཀྱུ, and སྐྱུ.

7) A horizontal combination of several consonants, but there is no prefix consonant or suffix consonant according to the restriction rules of native Tibetan Standard orthography to these positions. Examples of such combinations are ཀལཀལ, ཀངས, ཀལལ, and ཁངངས.

8) A horizontal combination of a consonant and a pre-composed character, but the consonant is neither the prefix consonant nor the suffix consonant. Examples of such combinations are ཀསི, ཀཆ, ཀཆི, ཀཀ, and ངར.

9) A horizontal combination of several pre-composed characters, but the last one is none of འི, འོ, and ཨ. Examples of such combinations are ཀིཀི, ཀིཁྀ, and ཀུཀུ.

3 The General Structure of Transliterating Characters

The collation of a transliterating character is not decided by its component letters directly. A transliterating character may be decomposed into several syllables firstly and then its collation is decided by those syllable series. Therefore, it is necessary to describe the syllable of transliterating characters.

3.1 The Collation Rules of the Transliterating Tibetan Dictionary

A transliterating character may be the vertical composition of basic consonant, foot consonant, and vowel and there are no concepts of prefix consonant, suffix consonant, and superscript consonant. Therefore, the phrases གཏི, མཐུར, and བཙན belong to the chapters ག, མ, and ད respectively; and the phrases ཀཆ, ཤྲས, and སཁྲ belong to the chapters ར, ལ, and ས respectively. Furthermore, a transliterating syllable may have two foot consonants and two vowels. For example, the syllable ཀྱྀ has two vowels. The first one is ྀ

and the second one is ◌ཻ. The diacritics ◌ི, ◌ུ, ◌ྲྀ, ◌ླྀ, ◌ཷ, ◌ཹ, ◌ཾ, and ◌ཿ should be treated as the first vowel either.

Before collation, a transliterating character or phrase should be decomposed into syllable series. For example, to collate the phrase ཀཻར, it should first be decomposed into syllable series ཀ་ཀ་ཻ་ར and then the collation of the phrase ཀཻར is decided by the corresponding syllable series ཀ་ཀ་ཻ་ར.

The collation of the single transliterating syllable is as follows.

1) The syllables with ཀ as the basic consonant are sorted as
ཀ་ཀཱ་ཀཱཻ ཀེ་ཀཱེ་ཀེཻ ཀཻ་ཀཱཻ་ཀཻཻ ཀུ་ཀཱུ་ཀཱཻུ ཀཱུ་ཀཱཱུ་ཀཱཱཻུ ཀྲི་ཀྲཱི་ཀྲཱིཻ ཀླི་ཀླཱི་ཀླཱིཻ ཀླཱི་ཀླཱཱི་ཀླཱཱིཻ ཀེ་ཀཻ་ཀེཻ ཀཻ་ཀཻ་ཀཻཻ ཀོ་ཀཽ་ཀོཻ ཀཽ་ཀཽཻ་ཀཽཻཻ (followed with those syllables that are the vertical combination of ཡ, ར, and ཀ with the vowels respectively).

2) The syllables with ཁ as the basic consonant are sorted as 1).

……

3) The syllables with ཨ as the basic consonant are sorted as 1).

3.2 The General structure of All Transliterating Characters

As mentioned above, a transliterating syllable is a pre-composition of a basic consonant with no more than two foot consonants and no more than two vowels. So, it has a general structure as shown in Fig. 3.

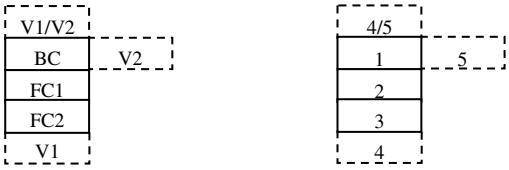

Fig. 3. The general structure of a transliterating syllable (left) and its sort order (right). Where V stands for the vowel, BC stands for the basic consonant, and FC stands for the foot consonant.

If a transliterating character cannot be represented by the general structure, it should be further decomposed into several syllables so that each of them can be represented by the general structure. For example, the character ཀྱྱཻ cannot be represented by the general structure directly but it can be decomposed into syllable series ཀ་ཀྱྱ, and both the syllables ཀ and ཀྱྱ can be represented by the general structure.

4 Collation of Transliterating Characters

The transliterating characters have two kinds collation: 1) collated with the rules of the native Tibetan dictionary and 2) collated with the rules of the transliterating character dictionary.

4.1 Collated with the Rules of the Transliterating Character Dictionary

When two transliterating characters are collated with the rules of the transliterating character dictionary, the scheme of the transliterating character collation consists of the following five steps as shown in Fig. 4.

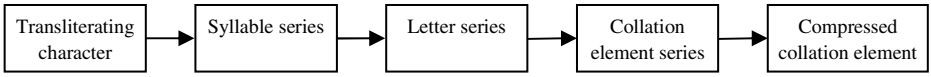

Fig. 4. The scheme of the transliterating character collation

Step 1: Decompose each transliterating character into syllable series first.

Step 2: Expand each syllable further into the letter series according to the sort order shown in Fig. 3. If there is no letter in a certain position, a space '□' is used instead.

Step 3: Replace each letter in the letter series with the corresponding collation element.

Step 4: Compress the collation element series.

Step 5: Compare the two compressed collation element series and we have got the collation result of two transliterating characters.

However, this paper just focuses on the first three steps.

To collate the two characters ཀྵ and ཀྴ, for example, they should firstly be expanded into syllable series 'ཀཥ' and 'ཀཤ' respectively; and then each syllable is further expanded into letter series, thus we have got 'ཀ□□□□ཀ ཥ□□□' and 'ཀ□□□□ཀ ཤ ཥ□□'; Finally, compare the two letter series as we compare two English strings and we have got the collation result of two characters ཀྵ and ཀྴ. Table 1 gives some more examples of this kind collation.

Table 1. The collation of the transliterating characters with the rules of the transliterating character dictionary

Characters	Syllable series	Letter series
ཀྵ	ཀཥ	ཀ□□□□ཀ ཥ□□□
ཀྴ	ཀཤ	ཀ□□□□ཀ ཤ ཥ□□
ཀྐ	ཀཀ	ཀ□□□□ཀ□□□□
ཀྃ	ཀཾ	ཀ□□□□ཨ ཱ□□ཾ
ཀཾ	ཀཾ	ཀ ཤ ཥ□ཾ
ཀྀ	ཨཱི	ཨ□□□□ཨ ཱ ཥ□ཾ
ཀྀཾ	ཀྀཾ	ཀྀ□□□□ཨ□□ཱི ཾ

4.2 Collated with the Rules of Native Tibetan Syllable Dictionary

A typical Tibetan syllable, such as བསྒྲིགས, is a two-dimensional combination of its letters. To syllable བསྒྲིགས, the letter ག at the center is the base consonant, the letter ས above the base consonant is the head consonant, the letter བ in the prefix position is the

prescript consonant, the letter ྲ, a variant of the letter ར, is the foot consonant, the sign ེ is a representation of the vowel ཨེ, and the two letters ག and ས after the base consonant are the postscript consonant and the post-postscript consonant respectively. The Tibetan orthography has strict restrictions to the letters in each position, for example, only the letters ག, ད, བ, མ, and འ can appear in the prescript position.

A few native Tibetan syllables have two foot consonants. For example, the syllable གྲྭ has two foot consonants ྲ and ྭ and the syllable ཕྱྭ has two foot consonants ྱ and ྭ. Furthermore, many transliterating characters, collected in the national standards of P. R. China on Tibetan Character Set, have two foot consonants.

When a transliterating character is collated with the rules of the native Tibetan syllable dictionary, a generalized structure should be constructed so that it can represent both the transliterating syllables and the native Tibetan syllables. The left part of Fig. 5 is such general structure.

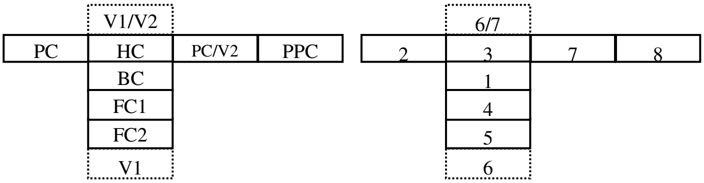

Fig. 5. The generalized syllable structure (left) and the sort order of the component letters (right). Where PC stands for the prefix consonant, HC stands for the head consonant, BC stands for the basic consonant, FC stands for the foot consonant, V stands for the vowel, PC stands fort the postscript consonant, and PPC stands for the post-postscript consonant.

When a transliterating character compares with a native Tibetan character or another transliterating character with the rules of the native Tibetan dictionary, it firstly should be decomposed into a serial of transliterating syllables as shown in the middle column of Table 2; and then each syllable is decomposed into a letter series by following the sort order shown in the right part of Fig. 5. If there is no letter in a certain position, a space '□' is used instead; Finally, compare the two letter series shown in the right column of Table 2.

Table 2. The collation of the transliterating characters and the native Tibetan characters

Characters or syllables	Syllable series	Letter series
ཀྐུ	ཀཀུ	ཀ□□□□□□□ཀ□□ུ□□
ཀྐྲུ	ཀཀྲུ	ཀ□□□□□□□ཀ□□ྲ□ུ□
ཀྵ	ཀཥ	ཀ□□□□□□□ཥ□□□□□
ཀྶྭཾ	ཀསྭཾ	ཀ□□□□□□□ས□□ྭ□□ཾ
ཀྲུཾ	ཀྲུཾ	ཀ□□ྲ ུ□ཾ□
བསྒྲིགས	བསྒྲིགས	ག བ ས ྲ □ ི ག ས
ཤྭུཾ	ཤཤྭུ	ཤ□□□□□□□ཤ□□ྭ
ཀྐྵྨྱོཾ	ཀྐམྱོ	ཀྐ□□□□□□□མ□□□□ོ

5 Conclusion

Compared with the native Tibetan characters, the transliterating characters are used not so popularly; however, there are more than six thousands of them. Therefore, it is necessary to study the collation of these transliterating characters. The paper proposes two structures that can deal with the two kinds of collation of transliterating characters: collated with rules of native Tibetan dictionaries and with the rules of transliterating dictionaries. Based on the proposed structures, all transliterating characters can be collated successfully and effectively with the rules of two different dictionaries.

Acknowledgment. This work is partially supported by NSFC under Grant No.60963016 and Key laboratory of Tibetan Information Processing, Ministry of Education of the People's Republic of China. The authors also thank the anonymous reviewers for their invaluable comments and suggestions.

References

1. An, S.: Sanskrit-Tibetan-Chinese dictionary. Nationalities Publishing House, Beijing (1991)
2. Zhang, Y.: Tibetan-Chinese Dictionary. Nationalities Publishing House, Beijing (1985)
3. Jiang, D., Kang, C.: The sorting mathematical model and algorithm of written Tibetan language. Chinese Journal of Computers 27(4), 524–529 (2004)
4. Huang, H., Da, F.: General Structure Based Collation of Tibetan Syllables. J. Inf. Comput. 6(5), 1693–1703 (2010)
5. Huang, H., Da, F.: Discussion on Collation of Tibetan Syllables. In: IALP 2010, pp. 35–38 (December 2010)
6. National Standard of PRC, Information Technology-Tibetan Coded Character Sets for Information Interchange-Extension A (GB/T 20542-2006). Standards Press of China, Beijing (May 2007)
7. National Standard of PRC, Information Technology-Tibetan Coded Character Sets for Information Interchange-Extension B (GB/T 22238-2008). Standards Press of China, Beijing (January 2009)