Collation of Transliterating Tibetan Characters

Heming HUANG[1,2*], 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

1 Introduction

Tibetan has two alphabets:

1）The native Tibetan alphabet used in daily life of Tibetan people and

The **native Tibetan** has 30 consonants and 4 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 4 vowels are ི, ུ, ེ, and ོ.

2）The transliterating Tibetan alphabet used specially to transliterate foreigner scripts especially the Sanskrit.

The **transliterating alphabet** has 34 consonants and 16 vowels. The 34 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, ཨ, and the 16 vowels are ཨ, ཨཱ, ཨི, ཨཱི, ཨུ, ཨཱུ, ཨྲྀ, ཨཱྲྀ, ཨླྀ, ཨཱླྀ, ཨེ, ཨཻ, ཨོ, ཨཽ, ཨཾ, and ཨཿ.

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

As shown in Fig. 1, the two alphabets are different but they share 28 letters ཀ, ཁ, ག, ང, ཅ, ཇ, ད, ན, ཏ, ཐ, ད, ན, པ, ཕ, བ, མ, ཡ, ར, ལ, ཤ, ཧ, ས, ཧ, ཨ, ཨི, ཨུ, ཨེ, and ཨོ in common.

The transliterating Tibetan has two kinds of collation.

1) All the characters need to be collated are just the transliterating characters and the collation of these characters follows the sorting rules of the transliterating Tibetan dictionary.

2) The collation of both the transliterating characters and the native Tibetan syllables, and the collation follows 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 (part of F300-F8FF, 0F0000-0F1645) 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 distinguishes 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.

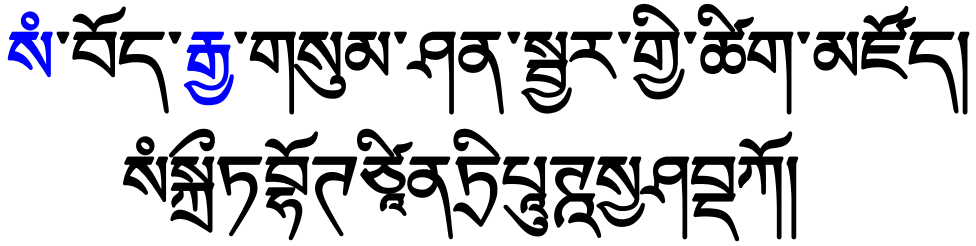


Fig. 2. A native Tibetan sentence (row 1) versus a transliterated sentence (row 2).

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 ◌྄, ྅, ◌྆, ◌྇, ྈ, ྉ, ྊ, or ྋ.

3) A pre-composed character has the transliterating consonants གྷ, ཛྷ, ཊ, ཋ, ཌ, ཌྷ, ཎ, དྷ, བྷ, ཥ, or ཀྵ.

4) A pre-composed character has 2 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 ཀྐེ, ཀྑ, ཀྥ, and ཏྟི.

5) A pre-composed character has 3 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 ཏྐྱ, ཏྐྐ, ཟྫྱེ, and དྡྭ.

6) A pre-composed character has at least 4 consonants. Examples of such characters are ཏྫྫྱ, ཏྡྱ, and ཏྫྫྱྱ.

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 གངྒམ, གལཀལ, གལལ, and ཁཊཊས.

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 གལི, གནྡ, གཉྩ, གཙྪ, and ངར.

9) A horizontal combination of more than one pre-composed characters, but the last one is none of ེ, ོ, and ྭ. Examples of such combinations are གེགེ, གེཚོ, and གུགྷ.

3 The general structure of transliterating characters

A transliterating character is vertical composition of a basic consonant, no more than 2 consonants, and no more than 2 vowels and there are no concepts of prefix consonant, suffix consonant, and superscript consonant.

The collation of the single transliterating syllable is as follows.

1) The syllables with ཀ as the basic consonant are sorted as

ཀ་ཀཱ་ཀི ཀཱི་ཀུ་ཀཱུ ཀྲུ་ཀྲཱུ་ཀླུ ཀླཱུ་ཀྱུ་ཀྱཱུ ཀྭུ་ཀྭཱུ་ཀྲྀ ཀཷ་ཀླྀ་ཀླཷ ཀེ་ཀཻ་ཀོ ཀཽ་ཀཾ་ཀཿ ཀྃ་ཀ྄ (followed with those syllables that are the vertical combination of ཀྱ, ཀྲ, and ཀྭ with the vowels respectively).

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

……

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

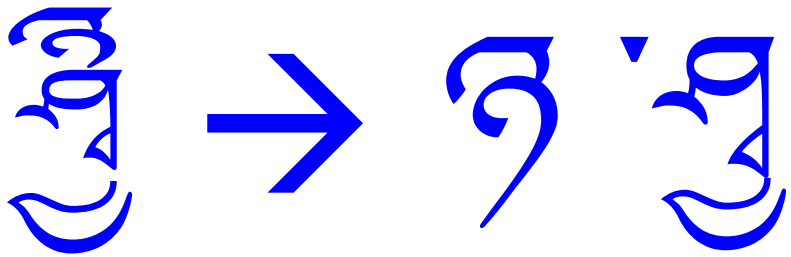
3.1 The general structure of all transliterating characters

A transliterating syllable is a pre-composition of a basic consonant with no more than 2 foot-consonants and no more than 2 vowels. Therefore, 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 syllable series so that each of them can be represented by the general structure.

For example,



4 Collation of the 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

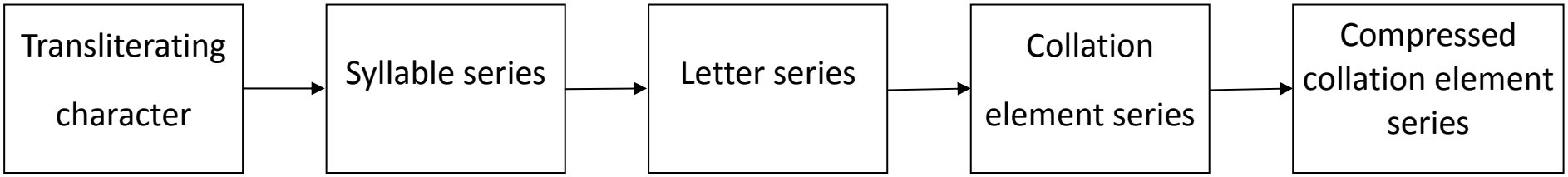


Fig. 4. The scheme of the transliterating character collation.

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.

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.

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

Characters	Syllable series	Letter series
ཀྱ	ཀཀྱ	ཀ□□□□ཀ ྱ□□□
ཀཱ	ཀཀཱ	ཀ□□□□ཀ ཱ ◌□□
ཆུ	ཀཆ	ཀ□□□□ཆ□□□□
སྤྲོ	ཀསྤྲོ	ཀ□□□□ས ྤྲ□□ོ
ཁྲོ	ཁྲོ	ཁ ྲ ◌□ ོ
ཀྲོ	ཀཀྲོ	ཀ□□□□ཀ ྲ ◌□ ོ
སྐྱོ	སྐྱོ	ས□□□□ཀ□□ྱ ◌ོ

4.2 Collated with the rules of native Tibetan syllable dictionary

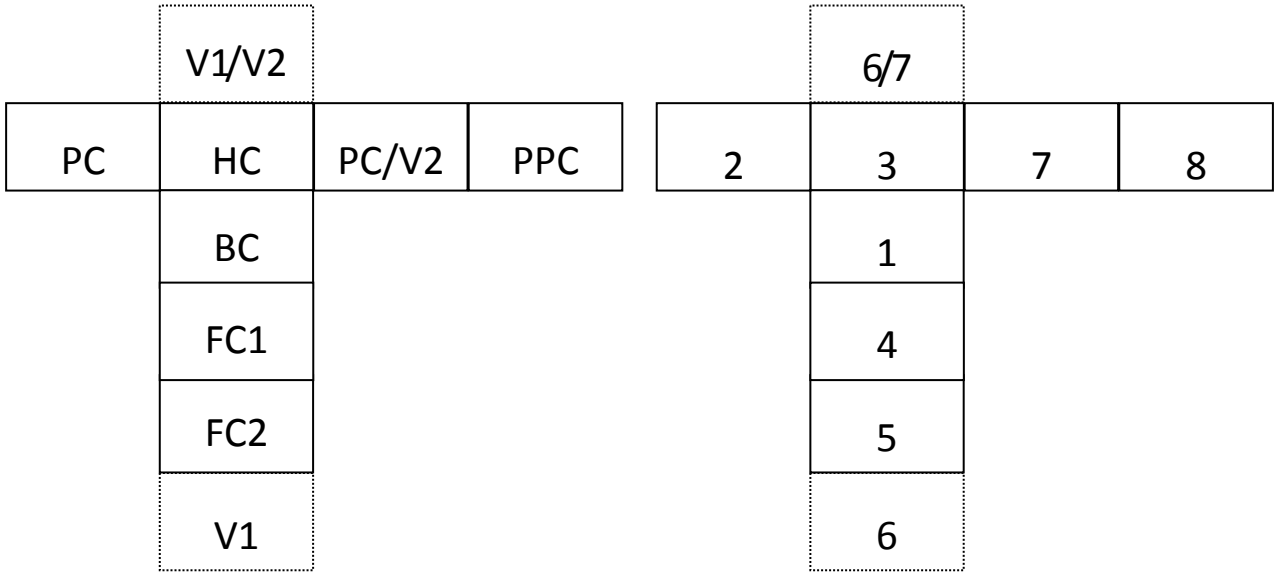


Fig. 5. The generalized syllable structure (left) and the sort order of the component letters (right). Where PC stands for the prefix consonant, HC stands for the head consonant, BC stands for the basic consonant, FC stands for the foot consonant, V stands for the vowel, PC stands fort the postscript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

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

Characters or syllables	Syllable series	Letter series
ཀཀ	ཀཀ	ཀ□□□□□□□ཀ□□ྐ□□□□
ཀཁ	ཀཁ	ཀ□□□□□□□ཀ□□ཱ□ྑ□□
ཀཆ	ཀཆ	ཀ□□□□□□□ཆ□□□□□□□
ཀཇོ	ཀཇོ	ཀ□□□□□□□ཇ□□ྒ□□ོ□
ཀཏོ	ཀཏོ	ཀ□□ཱ ྑ□ོ□
བཀྲིགས	བཀྲིགས	ཀབཇ ཱ□ི ཀ ཇ
ཀྵཱོ	ཀྵོ	ཀ□□□□□□□ཤ□□ྒ ྑ□ོ□
སྐྱིཿ	སྐྱིཿ	སྐ□□□□□□□མ□□□□ྱི ོ□

5 Conclusions

1) it is necessary to study the collation of these transliterating characters. Compared with the native Tibetan characters, the transliterating characters are used not so popularly; however, there are more than six thousands of them. Therefore,

2) 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.

3) Based on the proposed structures, all transliterating characters can be collated successfully and effectively with the rules of two different dictionaries.

Thanks !

Questions and suggestions?