View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Waseda University Repository

The Current Status of Sorting Order of Tibetan Dictionaries and Standardization*

Di Jiang

Academy of humanities, Shanghai Normal University, Shanghai, 200234
Institute of Ethnology & Anthropology, CASS, Beijing, 100081
jiangdi@cass.org.cn

Abstract. This paper discusses the problem of sorting orders of Tibetan dictionaries and Tibetan electronic databases. The alphabetical sequence of Tibetan language has being gradually formed by popular usage in the long history, in which many cognitive senses and cultural connotations lie embedded of Tibetan people, also there are a few influence of foreign elements and irrational elements. However, on the bases of the analyses of the background, the paper proposes three standardizing principles for compiling Tibetan dictionaries: agreement, compatibility, and rationality. In light of the three principles, the paper designs a set of digital codes for each letter or character and assigns distinctive sorting values to all existing words in the electronic dictionary with corresponding algorithm, which revises some serious errors in a previous paper. The method of compiling Tibetan word order mentioned above has been accomplished in our software system.

Keywords. sorting order Tibetan dictionary sorting values standardization

1 The Background

Although written Tibetan is a kind of phonemic language, its sorting order, in a dictionary or in an electronic database, is still not a simple issue. [1] The alphabet sequence of a language always shows by the word sequence of its dictionary. A dictionary sequence is not innate from the beginning, but gradually formed through common usage in a long history. So before discussing a compiling word order of a Tibetan dictionary, it is necessary to find out the development of its order, which may help readers to comprehend where the problem of the Tibetan alphabet order is.

Firstly, to let the issue clear, it is necessary to introduce briefly the Tibetan graphic structure. In figure 1, a Tibetan word or morpheme consists of not more than 7 letters, in which the base letter(Ba) is the core letter and the first one in structural sorting order, then the rest of the order is prefixed letters(Pr), head letters(Up), subjoined letters(Lw), vowels(V), suffixed letters(Sx). However, except the base letters, any other letters may be absent in the structure. Therefore, any structures with vacant letters will make special Tibetan graphic structures, which will change the order of different graphic forms (Figure 2). Furthermore, there exists different letter number in each position of graphic structures and the letters are in the order of alphabet sequence of their own.[2] The process of compiling order of a graphic form is that, choose letters by the alphabet sequence within the lowest position, such as position 7, after alternating all the letters in the layer, then enter into the above layer, and do the alternation again, after each change, return to the lowest layer, and run the previous circle. After the alternation goes through each position and each layer until the base letter changes, then a sorting order shifts to next graphic form with another base letter by the alphabet sequence. Read the complicated process in detail with reference [1] and [2].

* Funds Supported: Ministry of Education of China (MZ115-020), National Natural Science Foundation of China (No. 60173024, 60473135).

[1] There are only base letters and some subjoined letters in Unicode Standard, which are not available to process the word order directly.

[2] Thonmi Sambhota, the creator of written Tibetan scripts, pointed out in his great book Sum cu pa (On grammar) that the thirty letters are divided into seven groups and a half by four together in each group, which shows the letter sequence 1300 years ago.

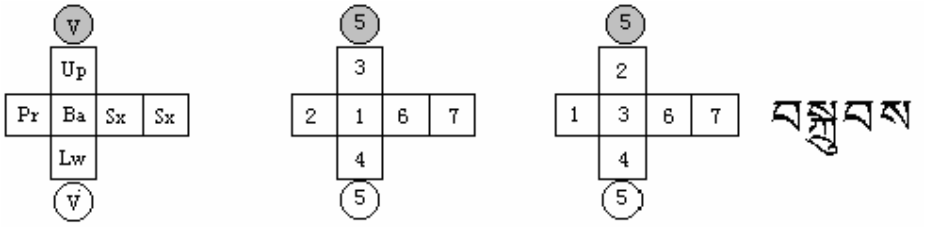

Figure 1 Graphic Structure Figure 2 Sequence of structure Figure 3 Sequence of handwriting an example

Csoma de Körös compiled the earliest Tibetan dictionary with an alphabet sequence in 1834, which is Essay towards a dictionary, Tibetan and English. [3] This dictionary is not really a dictionary with Tibetan structural and alphabet order in that he compiles it with Latin alphabet order, in which any prefixed letter and head letter are viewed as the same as basic letters. Hence, after compiling words with prefixed letters, such as g, words following are ones with head letter g-, no matter what the basic letters are. While the other base letters emerge afterwards words with prefixed g- will not appear any more in this way. It is the same with head letters. Once you compile base letter r, including head letter r- as well, it means that there is no head letter r- again in the rest of the dictionary. By the way, we also need to know that the Tibetan sorting order is quite different from its handwriting sequence (Figure 3).

In 1866, Hermann August Jäschke published his Romanized Tibetan and English Dictionary, which follows the tradition of Tibetan alphabet and the structural sorting order, which becomes the sorting model of Tibetan dictionaries later.[4] After Jäschke, another three important dictionaries came off the press. They are *Tibetan-English Dictionary with Sanskrit Synonyms* (by Sarat Chandra Das in 1902), [5] *A Tibetan Dictionary* by Dge bshes chus kyi grags pas in 1949, [6] *A Tibetan-Chinese Dictionary* by Zhang Yisun (chief editor) in 1985. [7] However, along with words growing in number, many errors and chaos emerge in sorting orders of Tibetan dictionaries, which will be discussed at next section.

2 Disagreements among Different Dictionaries

There are three reasons for the disagreement among different dictionaries. One is that there are no criteria on sorting orders of Tibetan-transliterated letters from Sanskrit. Another is that the nature of some character forms are still not clear, which causes compilers disagree with each other. The third is whether a dictionary ought to contain those forms of Tibetan sounds newly-emerged in its phonology or not.

The most troublesome problem comes from Tibetan-transliterated letters from Sanskrit. In such letters retroflex letters exercise much influence upon the sorting order, which are ཊ(tt), ཋ(tth), ཌ(dd), ཎ (nn), ཥ(ssh), and they appear with the sequence in the following dictionaries.

H.Ä.Jaschke views the retroflex letters as the same positions and layers as base letters in his A Tibetan-English Dictionary, however he only collects two of them, and gives the order like ཊ(tt), ཌ(dd), ཏ(t), ཐ(tth). [8] That is to say, the Tibetan-transliterated letters emerge prior to Tibetan letters in the sequence of dictionary entries. This method will surely do great damage to the principles of Tibetan tradition of 30 letters. [3] Therefore, it is not a good idea and method.

Melvyn C. Goldstein puts retroflex letters under the corresponding letters, such as entries with ཊ(tt) embedded in the entries with ཏ(t), ཌ(dd) in those with ད(d), in his Tibetan-English Dictionary of Modern Tibetan.[9] However, some disorders still exist in his dictionary. For example, under the condition of the

[3] About the Latin transliteration of Tibetan letters, please consult [8].

same vowels, ཊ(tta) appears after ཏ(ta). The sequence is ཏ(ta), ཊ(tta), ཏཱ(taa), ཏག(tag)……, ཏི(ti)……, ཏུ(tu). Yet ཌ(dd) changes the criterion, in which the sequence is ད(da), དག(dag), དལ(dal), ཌ(dda), ཌག(ddag), ཌི(di), ……, ཌུ(du), ……, དེ(de), ……, དེ(des), ཌེ(dde), ཌེན(dden), དོ(do), དོག(dog), ཌོ(ddo), ཌོག(ddog). In another words, ཌ(dd) emerges after the alternation of suffixed letters or before the alternation of vowels. Obviously, Goldstein makes wrong use of two different criteria on his compiling word order.

The compiling method of Zhang Yisun(chief editor) in A Tibetan-Chinese Dictionary is not the same as that of Goldstein for ཏ(ta), his sequence is as: ཏ(ta)…ཏི(ti)…ཏུ(tu)…ཏོ(to)…ཏོས(tos)…ཊ(tta)…ཊམ(ttam)…ཊི(tti)…ཏྭ(twa)…, namely ཊ(tta) follows ཏ(ta) with all the alternations of vowels and suffixed letters, or emerges only before subjoined letters. As for ཌ(dd), it seems to insert itself into different vowels, which is in the same sequence as that in Goldstein's dictionary.

Another type of letters, which may influence the sorting order, is Tibetan-transliterated composed characters from original monographic letters of Sanskrit. They are ཁྷ(gh), ཛྷ(ddh), ཐྷ(dh), ཕྷ(bh), ཛྙ(dzny), ཛྷ(dzh). However, the processing method for these letters in modern dictionaries always disintegrates them into two independent letters, the above one is base letter, and the other is subjoined letter. As subjoined letters, ཉ(ny) appears before the normal subjoined letter ྭ(w) and ྱ(y), ཧ(h) follows subjoined letters ྲ(r) and ླ(l).

As mentioned above, people have different ideas on the nature of the identification of the mark ཱ(va chung). It is a mark concerning three vowels, ཱ(aa), ཱི(ii) and ཱུ(uu). H.Ä.Jaschke did not give the mark a clear and definite identification, and let it appear freely in his dictionary, which is a common phenomenon in early dictionaries. In *A Tibetan-Chinese Dictionary*, ཱ (va chung) follows the corresponding vowels, and before the alternation between vowels and suffixed letters. For example, ཀ (ka), ཀཱ(kaa), ཀག(kag), ད(da), དཱ(daa), དག(dag); དི(di), དཱི(dii), དིག(dig). In *A Tibetan Dictionary* by Dge bshes chus kyi grags pas, ཱ(va chung) appears at the position before vowel alternations and after the alternations of suffixed letters. For instance, ཀ(ka), ཀག(kag), ཀར(kar), ཀཱ(kaa), ཀི(ki); ཤ(sha), ཤག(shag), ཤས(shas), ཤཱ(shaa), ཤི(shi), ཤུ(shu).

If mark ཱ(va chung) is a long vowel, is it a variant to vowels or another vowel, which is the nature of the mark. Different ideas will make different compiling word orders. Perhaps the process in *A Tibetan Dictionary* is a recommendable plan.

In recent years, some newly written forms emerged in Tibetan texts because of sound changing. Such as ཕྷ(hpha), which comes from the interpretation of sound /f/. How to arrange the new form in compiling word order in a dictionary is an issue of standardization. According to the experience of processing Tibetan-transliterated letters from Sanskrit, ཕ(ph) may be viewed as a subjoined letter, which is not in conflict with the traditional subjoined letters. Therefore, it may appear before subjoined letter *w* in line with the alphabet sequence. It is pity to say that it appears after the subjoined letter *w* in *A Dictionary of Tibetan-Chinese, Lhasa Dialect*.[10]

Besides the above problems, there are quite a few abnormal phenomena of sorting orders in dictionaries. Such as subjoined letter *w* emerges before the alternation of vowels and suffixed letters in *A Tibetan Dictionary*. For example, ག(ga), གྭ(gwa), གག(gag); གྲ(gra), གྲྭ(grwa), གྲག(grag); ད(da), དྭ(dwa), དག (dag), དང(dang), དྭང(dwang), ར(ra), རྭ(rwa), རག(rag). This means that all the structural sequence and layer based on base letters are in chaos or in confusion.

3 Standardization: Agreement, Compatibility, and Rationality in Sorting Order

It is necessary to set up principles in advance while compiling a dictionary. For Tibetan dictionary, there are three important things to do with. First, we need a strict standardization for the agreement accepted by the people through long social practice in history. Secondly, we ought to arrange forms, homogeny or heterogeneity, together rationally for compiling word order in common sequence. Thirdly, we should objectively understand the complicated phenomena, established by different usage, such as ambiguous symbols in semantics and symbolic variability.

In the light of the above three aspects of standardization, this session will give out a referential scheme to Tibetan dictionaries or electronic Tibetan dictionaries on its sorting order, in which some reasonable sorting values may be set up.

3.1 All 30 Tibetan letters will enter into the sorting order of dictionaries according to the traditional alphabet sequence, including retroflex letters, which follow the corresponding normal letters as subentries. See the sorting order in figure 1.

Table 1. the sorting order of Tibetan base letters

letter	ཀ	ཁ	ག	ང	ཅ	ཆ	ཇ	ཉ	ཏ (ཊ)	ཐ (ཋ)	ད (ཌ)	ན (ཎ)	པ	ཕ	བ
trans.	k	kh	g	ng	c	ch	j	ny	t(tt)	th(tth)	d(dd)	n(nn)	p	ph	b
value	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15
letter	མ	ཙ	ཚ	ཛ	ཝ	ཞ	ཟ	འ	ཡ	ར	ལ	ཤ (ཥ)	ས	ཧ	ཨ
trans.	m	ts	tsh	dz	w	zh	z	v	y	r	l	sh(ssh)	s	h	(a)
value	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30

As for the sorting values of retroflex letters, it is better to insert them before the alternation of vowels and suffixed letters, which adjusts to existing agreement, and may satisfy the requirement of compatibility and rationality. Here we need add an identification mark between subjoined letters and vowels to revise the useful value table in [1], in which *0* represents normal base letters, and *1* represents retroflex letters. see table 2.

Table 2. an example of sorting values of retroflex letters

	trans.	gloss	1st syl.							2nd syllable
			base	pre	up	sub.	mk	v	suf.	
ཏ་གུ་	ta ku	stick	09	00	00	00	0	06	00	0900000000600
ཊ་ཀ་	tta ka	annotation	09	00	00	00	1	01	00	0900000010100

3.2 Since many Tibetan-transliterated characters from Sanskrit can be head letters, with the sorting order changing above, to modify the sorting values in [1] is inevitable. See table 3.[4]

[4] Since a structural position may be absent in different syllabic structures, we pre-prepare *00* as a sorting value for the empty position at the beginning. The same is to the rest tables.

Table 3. head letters and their sorting values

letter		ཀ	ཁ	ག	ང	ཅ	ཆ	ཇ	ཉ	ཏ	ཐ	ད	ན	པ
trans.		k	kh	g	ng	ny	t	tt	th	tth	d	dd	n	nn
value	00	01	02	03	04	05	06	07	08	09	10	11	12	13
letter	ཕ	བ	མ	ཙ	ཚ	ཛ	ཝ	ཞ	ཟ	འ	ཡ	ར	ལ	ཤ
trans.	p	ph	b	m	ts	tsh	dz	w	r	l	sh	ssh	s	h
value	14	15	16	17	18	19	20	21	22	23	24	25	26	27

In [1] we have discussed the special situation on different types of syllable structures, in which an identification mark is used to sorting values of prefixed letters. Here we copy them down in table 4.

Table 4. prefixed letters and their sorting values

Letter		ག	ད	བ	མ	འ
Trans.		g	d	b	m	v
Value	00	01	02	03	04	05
Value with identification mark	10	11	12	13	14	15

3.3 While we disintegrate some of the typical composed characters from Sanskrit, new subjoined forms need to add into the subjoined table in [1], which include ྐྵ(-ny) from ྫྷ(dzny), ྷ(-h) from ྒྷ(gh), ྡྷ (ddh), ྡྷ(dh), ྦྷ(bh), ྫྷ(dzh), ྴ(-ssh) from ྐྵ(kssh), and reduplicated subjoined symbols ྈ(-hw), ྉ(-hr), and ྥ(-ph) from ྥྷ(hph) in Lhasa dialect. There are some other changes.[8] See the following table 5.[5]

Table 5. subjoined letters and their sorting values

letter		ྐ	ྑ	ྒྵ	ྥ	ྭ	ྮ	ྱ	ྲ	ྴ	ྵ	ྶ	ྷ	ྸ	ྐྵ	ྺ	ྻ
trans.		(k)q	(p)q	ny	ph	w	y	yw	yr	r	rw	ry	l	ssh	h	hw	hr
value	00	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16

3.4 The reason we revise the vowel table and their values comes from the further comprehension on the nature of ྭ(va-chung), which has been discussed above. To match the three vowels with va-chung, we regard ེ(ee) and ོ(oo) from Sanskrit as long vowels although they may be diphthongs ‘ai’ and ‘au’ in Sanskrit. The position of each long vowel is in the rear of the corresponding short vowel in the sorting sequence. In addition, some archaic retroflex vowels ྀ(·i) and syllabic vowels ཷ(<ཫ ཱི i), ཹ(<ཬ ཱི i), ཷ (<ཪ ཱི i), ཹ(<ཬ ཱི i) from Sanskrit may be included in the same category.

Table 6. vowels and their sorting values

letter	ཨ	ཱ	ི	ཱི	ྀ	ཱྀ	ུ	ཱུ	ེ	ཻ	ོ	ཽ
trans.	a	aa	i	ii	·i	·ii	u	uu	e	ee	o	oo
value	01	02	03	04	05	06	07	08	09	10	11	12

[5] As the revised number of subjoined letters is now more than 9 items, we expand their sorting values to two digits. In addition, we also add an identification mark to retroflex letters within base letters, therefore, the number of total sorting values ought to be 13 digits for Tibetan sorting order in electronic dictionaries.

3.5 The sorting order of suffixed letters is also under the standardization. That is to say, Tibetan-transliterated letters from Sanskrit always follow the corresponding Tibetan traditional letters. Moreover, some special symbols will find their positions according to customary usage.[8]

Table 7. suffixed letters and their sorting values

letter		ཿ	ཀ	ཁ	ག	གས	ང	ངས	ཉ	ཏ	ཊ	ཐ	ཋ	ད	ཌ	ན	ནད	ཎ
trans.		hq	k	kh	g	gs	ng	ngs	ny	t	tt	th	tth	d	dd	n	nd	nn
value	00	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17
letter	པ	ཕ	བ	བས	མ	མས	ཾ	ཙ	ཚ	ཛ	ཝ	འ	འང	འམ	འན	འས	འི	འིམ
trans.	p	ph	b	bs	m	ms	mq	ts	tsh	dz	w	v	vng	vm	vn	vs	vi	vim
value	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35
letter	འུ	འུའི	འུར	འུས	འེ	འོ	ཡ	ར	རད	ལ	ལད	ཤ	ཥ	ས	ཧ			
trans.	vu	vuvi	vur	vus	ve	vo	y	r	rd	l	ld	sh	ssh	s	h			
value	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50			

4 Conclusion

The basic functions of a dictionary include knowledge, sciences, and retrieval. The above discussion sums up the sorting orders of Tibetan words in some dictionaries, revises some errors in [1] according to the standardized principles of compiling dictionaries. Now on the bases of the further comprehension, we adjust all the elements and algorithms to Tibetan electronic dictionaries, and lay a solid foundation for Tibetan language processing in future.

Reference

1. Jiang, Di, Kang, Caijun: The Sorting Mathematical Model and Algorithm of Written Tibetan Language. *Journal of Computer Science and Technology*. (2004) Vol. 27: 4. 524-529
2. Jiang, Di, Zhou, Jiwen: On the Sequence of Tibetan Words and the Method of Making Sequence, *Journal of Chinese information processing*. (2000) Vol. 14: 1, 56-64. also in: Hiroyoshi Ohara(ed.): Collections of International Conference on Multilingual Text Processing'98. Waseda University. Tokyo, Japan (1998) 9-20
3. Csoma de Körös, Alexabder: *Essay towards a dictionary: Tibetan and English. Prepared with the Assistance of Bandé Sangs-Rhyas Phun-Tshogs, a learned Lama of Zangskár*. Culcutta (1834). [reprinted Budapest: Akadémiai Kiadó, 1984]
4. Jaschke, H.Ä.: *A Tibetan-English Dictionary*. Delhi: Motilal Banarsidass (1881). Reprinted, London, 1980
5. Das, Sarat Chandra: *Tibetan-English Dictionary with Sanskrit Synonyms*. Calcutta, India (1902)
6. Dge bshes chus kyi grags pas: *Tibetan Dictionary*, (xylography, 1949). Reprinted in Beijing: Nationalities Press (1957)
7. Zhang, Yisun (ed.): *A Tibetan-Chinese Dictionary*. Beijing: Nationalities Press (1985)
8. Jiang, Di: Approaches on Methods of the Latin Transliteration in Tibetan, *Minority Languages of China* (2006) vol. 1, 45-53
9. Goldstein, Melvyn C.: *Tibetan-English Dictionary of Modern Tibetan*. New Dlhi: Rakesh Press (1975), reprinted in 1978
10. Yu, Daoquan (ed.): *A Dictionary of Tibetan-Chinese, Lhasa Dialect*. Beijing: Nationalities Press (1983)