A Fast Input Method for Tibetan Based on Word in Unicode

Weilan Wang, Lingwang Kun

Abstract—**The database is essential resources for an intelligent Tibetan language inputting system in Unicode. This article mentions Tibetan language input system database including syllable and vocabulary. The database of syllable and vocabulary are established by the code rule of Latin Transcribing and root Latin Transcribing, respectively, and the frequency of words are the statistics based on the Tibetan language corpus of 60M. Searching possible syllables and words based on inputted Latin Transcribing string or root Latin Transcribing string; matching the template syllables and words according to the arithmetic of optimal match, the key assignments of input code can be able to obtain. Then the key assignment point to the linked list of the syllables or words corresponding, and same code syllables or words are sent to the candidate windows in the linked list, thereby inputting of searches of syllables, Sanskrits and words are implemented. So far the input method of Tibetan language is a very fast and efficient in Unicode.**

Key words: **Syllable Database, Vocabulary Database, Fast Input, Tibetan language Unicode**

I. Introduction

In order to user use computers effectively, one must be able to quickly and accurately input one's words into a computer, so it is necessary to research input methods of Tibetan word; what is the best input encoding that we are very interested in Tibetan inputting? Tibetan language is a sound-spelling languages, there are 30 consonants and 4 vowels through rules to spell from left to right and above to below form syllables which are unit to compose words, and we have explained the reason of use Latin Transcribing of Tibetan language as input encoding [1]; Latin Transcribing of Tibetan are organized such that spaces are to be placed between syllables of words. Our software system that we have finished allows users to use English keyboard to type in the Latin Transcribing of Tibetan words or sentences, we just type Latin Transcribing of syllables, words and sentences of Tibetan and implement Tibetan inputting. Those are our foregone work. The paper [2] "propose a Tibetan input method based on Tibetan syllable" for mobile phone. Our research will solve the issue of Tibetan fast input in computer and they include syllable, Sanskri and word in Unicode.

This paper is a farther study and development on Tibetan language input method in Unicode, at some times, we also finished the input to put together with syllable, Sanskrit, and word, and do not switch when input syllables, Sanskrits, and words. This paper is organized as follows: in section 2, the database of syllable and vocabulary and their structure of storage are discussed. A search inputting of syllables and Sanskrits, words are introduced in section 3. Section 4 is encoding evaluation and result analysis. Some experimental results are reported in section 5, conclusion for farther research.

Manuscript received October 20, 2007. The research supported by China Scholarship Council, The Tackle Key Problems in Science and Technology Foundation of Gansu Province under Grant (NO: 2GS064-A52–035-04) and the Scientific Research Foundation for the Returned Overseas Chinese Scholars, Ministry of Personnel of People's Republic of China.

Weilan Wang is with the Information Technology Institute, Northwest University for Nationalities, Llanzhou, Gansu 730030 China (phone: 086-13609385804; e-mail: wangweilan@ xbmu.edu.cn).

Lingwang Kun is with the Gansu Meteorological Information & Technical Support & Equipment Center, Data Laboratory, Lanzhou, Gansu, 730020, China (e-mail: klw03@163.com)

II. The Database of Tibetan Input

A. Database of Syllable, Sanskrit and Structure of Storage

To design an application and intelligent Tibetan language input software, a syllable table of standardization is prerequisite. But how many syllables in the Tibetan practical? How about their frequency distributing? Up to now we have not an authoritative last word, so these works begin from the statistic of canonical syllables based on the Tibetan language corpus. Our previous work [3] [4] analyzed and discussed the statistic results and situation of frequency distributing about the Tibetan language syllables. We also design the arithmetic according to spelling rules of syllable, and generate the Tibetan syllables. The single-character, double-character, triple-character, and quadruple-character syllables are 445, 4985, 7122 and 2237 respectively, so the total numbers of syllable are 14789. The previous work demonstrated in this aspect and they are some very significant results. We also got the Sanskrits which are 5280. 5 reverses and 6 vowels (ཪ called long vowel) of the Sanskrit are encoded in the table 1 as follow.

Table 1. Sanskrit Encoding

ད	ན	ཤ	ཐ	ཏ	ེ	ི	ཾ	ོ	ཿ	ཨ
Da	Na	SHa	THa	Ta	E	I	M	O	H	A

The Sanskrit is a vertically "stacked" by 30 consonants, 4 vowels, 5 reverses and 6 vowels of Sanskrit, the code of 30 consonants and 4 vowels was dictate in[1].

The rules of encoding of the Sanskrit are: if a Sanskrit take a vowel, its encoding is from top to bottom every letter's code but omit "a", the last letter of the code is Latin

Transcribing of the vowel. Such as: ཀྭོ=>*kwO* , ཉྷེ =>*NTHe*; if a Sanskrit take 2 vowels, its encoding like the code of one vowel and the order of two vowels is from left to right, or from bottom to top such as: ཀྵྱུཾ=> *kSHyuM*, ཀླིཾ=> *kliM*, མྦྷཾ => *mOM*, ངྦྷཱི => *ngbhAi* ; if a Sanskrit do not take vowel, its encoding is still from top to bottom and the letters do not take "a" except the last letter such as: ཛྙ => *dzNa*, ཛྦ=> *dzba*, ཛླ=>*dzhla*.

Collection of the syllables and Sanskrits form a list, according to the rules of encoding, generation the encoding list of syllables and Sanskrits in Unicode. The database of the syllables and Sanskrits is expressed through some operations such as classification, taxis and so on, in two linked list structures. One is the linked list structure of encoding of the syllable and Sanskrit:

```
typedef struct SyllableCode {
WORD wKey;
TCHAR szSY[MAX_ SYLLABLECODE _LEN];
struct  SyllableCode  FAR *SyllableNext;
} SyllableCode;
```

Where wKey expresses the value of syllable or Sanskrit, and their data type is WORD; szSY[MAX_SYLLABLECODE_LEN] is the groups number of syllables or Sanskrits encoding , and their data type is TCHAR, each element in the array save one the encoding corresponding all syllables or Sanskrits; MAX_ SYLLABLECODE _LEN expresses the maxim code length of syllables or Sanskrits.

The other is linked list structure of syllables or Sanskrits:

```
typedef struct Tibetan{
TCHAR szTibetan[MAX_ TIBETAN _LEN];
struct Tibetan FAR *TibetanNext;
} Tibetan;
```

Where szTibetan[MAX_TIBETAN_LEN] is groups number of syllables or Sanskrits, and the type is TCHAR; MAX_ TIBETAN _LEN is maxim length of the syllables or Sanskrits.

B. Vocabulary Table and System Dictionary

We have finished a Tibetan words list of 57909, the frequency of words is the statistics based on the Tibetan corpus of 60M. According to the rules of Tibetan Latin Transcribes, a vocabulary list coded can be produced by encoding list of syllables and Sanskrits in Unicode. Then the System Dictionary are designed and produced by the database of Syllable and Sanskrit.

The memory form of System Dictionary is:

Word, Key Value of Word, Length of a Word (the syllables number in a word), Attribute, Type

Where the *Attribute* indicates the frequency of the word, the *Type* indicates or User Dictionary (which store new word or phrase which are combined to use System Dictionary and syllable list by user on the same structure with the System Dictionary)

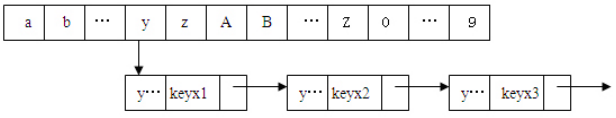
III. RETRIEVAL INPUT OF THE SYLLABE,SANSKRIT AND WORD

We proposed to take the syllable as the smallest code unit, the glossary input as primarily code input scheme, and it is more standard and fast than take character as the code unit's input method. Firstly, look at from the information theory angle, the Tibetan multi-dimensional entropy must be smaller than the one-dimensional entropy, the input method of the word, the phrase and the sentence compares with the character input method to need few input information. Next, say from the humanist angle, when keyboard entry, people's thought defers to the word, phrase or sentence of certain meaning. The syllable database and the input dictionary also are the template of the syllable or Sanskrit, word, phrase. The syllable or the word by input may through searches in fuzzy, high frequency to match the template in the syllable database or System Dictionary, to determine the possible the syllable or word as output result.

A. Retrieval Input of Syllable or Sanskrit

Searches input of syllable or Sanskrit is exemplified in Fig 1 and Fig 2. The order of the key value are: key1,key2,…,keym, m is number of syllable's key value. Each key value is correspondence at least one syllable, and one more syllables saved by the linked list. Because in the syllable database including all of syllables and 10 numerals of Tibetan language, first encoding letter of syllable is a, b, c,…,z separately in Figure 1, and Sanskrit also has the capital letter A,B,…,Z, 10 the Tibetan numerals of codes are 1, 2,…,9 separately. The judgment of first letter is a, b,…,z, or A,B,…,Z, or 0,1,…,9 according to the input code. Each input code may correspond one encoding key value, also may have many encoding key values, sorts by the letter order, the linked list form links. The first item is the input code, and the second item is the key value in the linked list structure. The keyx1, keyx2, keyx3, … , is one of the key1,key2,…,keym respectively.

Message retrieval process: Found first the most suited code with the input code, also has obtained the key value of this code; according to the key value corresponding the chain of Tibetan syllable, the same code syllable will be send to candidate window in the chain. Linked list simultaneously direct to next key value, the continuation of above the process, syllables by search in syllable database will be sent to candidate windows and user choices one of them every time. In this process, with the input string determined gradually, namely the key value correspond determination, syllables quantity which are sent into the candidate window will be reduction. Namely if input's information more many, the retrieval results will be more accurate.



Figl. Retrieval the key value according to the code

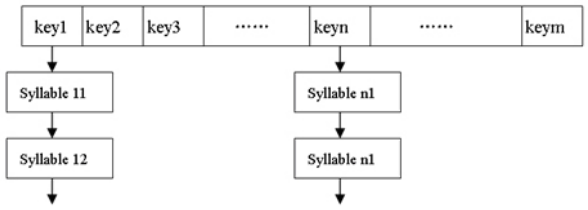


Fig 2. Retrieval the syllable according to the key value

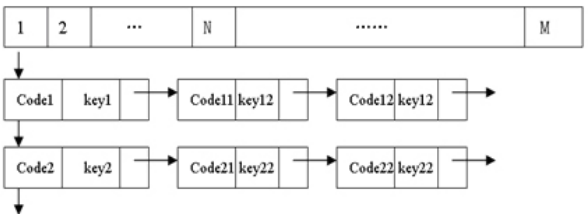


Fig 3. Retrieval the word

B. Retrieval Input of Vocabulary

There are System Dictionary and the user dictionary when user input vocabulary, the retrieval input way are completely same in the two dictionaries. We take the System Dictionary as the example, shown in Fig 3: 1,2,…,N,……,M is the key value of first syllable of word separately ,and max value is M; ode1,Code2,…,Code11,Code12,…,Code21,Code22, … , are input code of syllable , and the syllable's correspond key value is key1, key2,…, key11,key12,…,key21, key22, … separately. If the key value of first syllable is different, the linked list is downward zipper; towards right the zipper are expressed key value same to the word of the first syllable. According to the key value of second syllable again, if the key value is same with the key value of first syllable, the linked list will continue to pull toward right, until all syllables of the word are link in the chain.

At the same time, each Tibetan word is storage as following: dwBaseOffset + dwMapFileOffset + sizeof (SyllablePH). Where the dwBaseOffset is the start address of the word; the dwMapFileOffset is deposit the word's offset address; the SyllablePH is Tibetan attribute.

IV. CODE EVALUATION AND RESULT ANALYSIS

A. Encoding Efficiency

Tibetan language syllable as the minimal encoding unit, the input of Latin Transcribes transform into Tibetan, the encoding scheme is standard, simply and easy to hold. How about the encoding efficiency? 16744 (inclusion some especial syllables that postfix is ཞེ , ཅི , འུ , བོ for input convenience) syllables are divided into 23 (letters q,f and x not used in Latin transcribes) subsets according to each Latin Transcribes of syllable or root start, the syllables in each subset has same first letter of Latin Transcribes.

Table 1 show that Latin Transcribing of syllable have repeating code only while first letter is "g", the coding "gyib" corresponding syllables are: གཡིབ and གྱིབ, and the "gyibs" corresponding syllables are: གཡིབས and གྱིབས. Namely the suffix of those syllable is "*ga*" and root is "*ya*", or root is "*ya*" and root is "*ga*" subjoined is "*ya*", and the rest all of syllables have not repeating code, so it high efficiency in encoding. But numbers of maxim repeating code of Latin transcribing of root start are 14. So it is simple that Latin transcribing of syllable code and root start's code, especial Latin transcribing encoding is low repeating code, the number of possible choices corresponding to that Latin transcribing encoding is greatly reduced, and input can often be made much more efficient. On the other hand, we also design root-encoding way for people's thinking custom. Because the starting point of people thinking is root of each syllable of word, so the input way also shows more efficient when input words. For example, the encoding of word

"དོར་ཞི་གདོང་དྲུག་སྙྱེམས་པའི་བློ་ལྡན་" is:

dor zhi gdong drug snyems pavi blo ldan

and also can be type: "*d z g d s p o l*" which are 8 letters key and 7 spaces key. The word "ཤཀྵ་ཊ་ཡ་ན་བྱ་ཀ་ར་ཎ" has Sanskrit and its input encoding is :

sha kO Ta ya na byA ka ra Na

or only type:

s k T y n b k r N .

It is can renew combination while the input words not exist in System Dictionary. In order to input the word: རླངས་ཆེན་, for example, user can input Latin Transcribing of root start: *rlangs chen*, user choice first syllable from the candidate window, and then choice second syllable from the other candidate again, a new phrase are send into current document and newly formed word or phrase are added to user dictionary and can be used next time.

Table 1 Compare of repetition code between Latin Transcribes of syllable and root start

first letter	Number of syllable	repetition code	root letter	Number of syllable	repetition code
a	82		a	82	
b	3516		b	388	6
c	205		c	219	
d	1218		d	294	
e	1		e	1	
g	1137	2	g	397	14
h	261		h	265	2
i	1		i	1	
j	86		j	93	6
k	630		k	643	8
l	960		l	82	3
m	1582		m	290	4
n	256		n	386	8
o	1		o	1	
p	537		p	561	4
r	1523		r	174	2
s	2348		s	373	3
t	531		t	531	9
u	1		u	1	
v	1400		v	89	
w	86		w	174	
y	87		y	90	2
z	281		z	290	3

B. The Course of Encoding Input for Syllable and Word

Table 2 shows an input's example for syllable: དབྱངས, its input code is "*dbyangs*", the first column denote the course of typing, the second column is candidate syllable in candidate window.

Table 2. An example for syllable's input

Courese of encoding input	Syllable list in first candidate Window	Note
d	ད དབ དད དག དཀ དམ དན དང དཟ དས	all of ayllable of first coding is "d"
db	དབཅ དབཆས དབད དབག དབགས དབཀ དབམ དབན དབང དབངས	All of ayllable of previous two coding are "db"
dby	དབྱས དབྱཀས དབྱམས དབྱངས དབྱད དབྱེ དབྱེས དབྱིས དབྱུས དབྱོ	choice 3 in the candidate window can input the syllable of need
dbya	དབྱངས དབྱད དབྱམས དབྱཀས དབྱས	Only five syllables while coding is "dbya" and fourth is the syllable.
dbya ,dbyang, dbyangs	དབྱངས	Only one syllable corresponding the input of "dbyan", "dbyang" or "dbyangs".

Table 3 shows the input course of word: དབྱངས་ཅན་དགྱེས་གླུ , its input code is "*dbyangs can dgyes glu*", the first column denote the course of typing, the second column is candidate words in candidate window. Of course, if type " *d c d g*" , the word can be find in candidate window.

Table 3. An example for word's input

Courese of encoding input	Word list list in first candidate Window	Note
dbyangs c/ca /can	དབྱངས་ཅན	all of ayllable of first coding is "d"
dbyangs can dgyces	དབྱངས	All of ayllable of previous two coding are "db"
dbyangs can dgyes glu	དབྱངས་ཅན་དགྱེས་གླུ	choice 3 in the candidate window can input the syllable of need

V. CONCLUSION

The research of Tibetan language keyboard input method is essential to information processing, an ideal input method should be standard in Tibetan, easy operation, adaptable different user usage habit. We think our designed and realized the input method of the Tibetan syllable, Sanskrit and word in Unicode by searches the database of the syllable and dictionary and use standard keyboards, there are two way type Tibetan, one is Latin Transcribing encoding which is greatly reduced the number of repeating code, the other is root Latin Transcribing encoding which is to begin encoding from the root of a syllable, which is coincidence thinking habit. Our methods provide a rational and feasible input encoding scheme for Tibetan language input more fast and effective. Our further research include the statistic of the co-occurrence probability of Tibetan language words, so that the statistical results can be added the input method to make Tibetan language typing with associational function, namely from one word of inputting associate next words by the co-occurrence probability in the input process.

REFERENCES

[1] WANG Wei-lan. Intelligent Input Software of Tibetan. Computer Standards & Interfaces. 29 (2007) pp462–466.

[2] Sen Lin, Yuan Dong, Shengyuan Wang,Ting Li,Nymatrashi, Pudun.A Tibetan Input Method Based on Syllable Word for Mobile Phone.Proceeding of the Second International Conference on Embedded and Systems(ICESS'05)

[3] Wang Weilan. "The Frequency-Rank of Language Unit in Modern Tibetan " Science Technology and Engineering, May, 2004, pp413-417

[4] Wang Weilan and Chen Wanjun. "The Frequency and Information Entropy of Tibetan Character and Syllable" Terminology Standardization & Information Technology, June, 2004, pp27-31