Statistical Analysis of Words Based on Textbooks Database of Tibetan Primary School

Yonghong Li

Key Lab of China's National Linguistic Information Technology, Northwest University for Nationalities, Lanzhou 730000, China

LYHWEIWEI@126.com

Abstract

This paper based on Tibetan edition of Chinese textbook (from volume 1 to volume 12) in five northwestern provinces (autonomous regions) as research goal, and established corpus of Tibetan language textbook. The paper got statistics from optional, coverage rate of Tibetan which is commonly used to word's repetition rate with these 12 volumes Tibetan language, offered an integrated quantitative standard to textbook evaluation, and offered basis teaching materials for its development , the purpose is that new teaching materials can meet the needs of Tibetan basic education ,and even meet the needs of all Tibetan educational development.

Keywords

Tibetan, primary school textbooks, the statistical analysis, words.

1. Introduction

The teaching materials of Tibetan language in primary school are an important component of contemporary Tibetan basic education, which plays a vital role in universal compulsory education in Tibetan area. In order to make commonly used words in primary language teaching material can meet the needs primary's cognitive features , the design of teaching materials must based on a vocabulary outline, which should be a scientific system . With the rapid development of computer and introduction of large-scale corpus, meanwhile the introduction of statistical methods can be help design of Tibetan language teaching materials and make it to become more scientific and rationalization [1].

A number of expositions on Tibetan language teaching materials concerned in its content and teaching methods in primary schools, we can take some paper as examples , for example , a study of Tibetan language in Qinghai[2], the discussion on write Chinese textbooks in Tibetan primary and secondary schools[3].

The study of this paper based on a collaboration teaching materials the Tibetan language in primary school (from edition 1-12), which used in five provinces (autonomous regions), edited and published by Qinghai Nationalities Publishing House [4]. This paper has discussed many different statistical study from the selection of commonly used characters, the coverage rate of commonly used characters to the repetitive rate of characters.

2. Tibetan word

Tibetan language is an alphabetic writing developed on the basis of Sanskrit Tiancheng front type. It has 30 consonant letters and 4 vowel symbols among which /a/ is zero. Tibetan language has a strict and complete set of rules of letter combinations, and its characters flow is a two-dimensional show which is horizontally written from left to right. In accordance with the structure location of letters in syllables, traditional Tibetan grammar divides the letters into "basic character", "upper character" "sub character", "forward character", "post character", and "re-post character". The basic word is the nucleus of a whole Tibetan word and 30 consonant letters can all act as the basic word while vowels can not. Other parts can all be empty except basic words. A Tibetan word can consist of seven components (including vowels) at its maximum as is in figure 1.

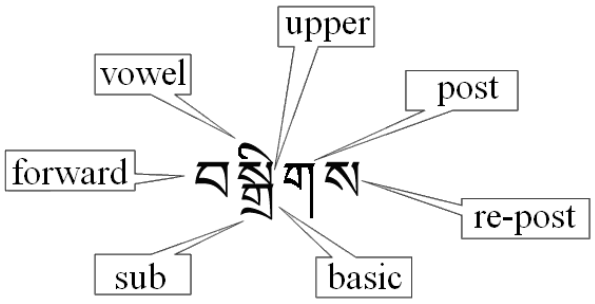

Fig. 1 The structure of Tibetan word

3. The Table Of Commonly Used Tibetan Characters

Teaching Tibetan language is a multiple element hierarchy system, it is very important that how to teach word for improving the efficiency of the Tibetan language teaching. Reading and writing literacy as prerequisite, and students should be master commonly used characters with high capacity in combine new words. There has not yet issue the commonly used words and secondary commonly used words in Tibetan. We organized and collected text corpus more than 10,000 sentences , which included Tibetan and Chinese dictionary, a total of 1,298,140 Tibetan word (between two-node), and we amounted 6323 Tibetan words which rule out repeat , their average frequency is 207.Tibetan characters were divided into three sections according its frequency distribution, ultra-high frequency words, high-frequency words and middle high-frequency words. High-frequency words and middle high-frequency words are commonly used words, middle-low frequency words are secondary commonly used words, and low–frequency words with ultra-low-frequency word are not commonly used (see table1).

Table 1 Division Table of Tibetan Word Frequency

Type	Section	Segment	Quantity	Frequency（%）
Commonly used words	ultra-high frequency words	1-100	100	53.27
	high frequency words	101-500	400	31.79
	middle high frequency words	501-1000	500	9.17
secondarily commonly used words	middle-frequency words	1001-1500	500	3.07
	middle-low frequency words	1501-2500	1000	1.93
rarely used words	low-frequency words	2501-4000	1500	0.59
	ultra-low frequency words	4001-6363	2362	0.18

4. The Coverage Frequency Of Commonly Used Words

The selection of teaching common words is one of the more difficult problems in teaching. We imputed 293 texts from textbook volum1 to 12 and, established a Tibetan language textbooks database of primary school, the total 121,484 words, and 10,593 sentences. Because the major target of first volume of textbook is to study letters, we mainly focus on 2-12 volumes , and divided new words of each text into two types , namely recognition of word (words that first appeared in the text , but not appeared in words list) and mastering word (words appeared in the worse list). The requirements of the recognition of word will be identified, read, and can be understand the meaning of the word. The requirement of mastering word is will be identified, read, write and master its pronunciation, shape and meaning. The words appeared in primary textbooks should cover at least 95 percent in the table of commonly used word. Some words, however, can hardly appear in the text, should be appear in the practice exercises of text. Table 2 shows coverage of commonly used words, secondary used words and rarely used words in both categories.

Table2 Commonly used words compared with the number of words appeared in texts

volume	NT	NWT	CWT	NCWT	NNWT	AWT	NWL
1	0	0	0	0	0	0	223
2	29	546	316	27	7	19	185
3	28	434	107	41	8	16	193
4	28	253	11	18	11	9	170
5	25	263	6	32	8	11	131
6	25	232	0	24	7	9	152
7	24	272	0	26	11	11	135
8	24	222	0	20	12	9	145
9	24	232	0	23	22	10	124
10	24	173	0	5	9	7	107
11	22	197	0	12	16	9	70
12	21	98	0	2	3	5	68
Total	275	2922	440	230	114	11	1703

NT :the number of test

NWT: he number of words appeared in text

CWT: the number of commonly used words appeared in text

NCWT: the number of secondarily used words appeared in text

NNWT: the number of non-common used words used words

AWT:average number of new words in each text

NWL: the number of words in new words list

According to the statistical analysis , the result of this table shows that this teaching materials requirement primary school students should recognize 2922 words , can be write 1703 words . By controlling the amount of writing to low –grade students, we can reduce the barriers of writing for young students, so that they can widen their amount of reading, expand their horizons, and absorb more knowledge. However, the gap between the amounts of recognized words and really can be write by Tibetan students is too large, this phenomenon especially evident in the lower-grades. It wills influence on actual teaching. Bilingual teaching is common in primary school of Tibetan area, thus students should master Tibetan and Chinese , text is more difficult in the view of words , which students should be recognize .

The trend of the amount of new words, coverage of commonly used words, and secondarily commonly used words is significantly decrease with increase of textbook volumes, it meets the requirement of recognize words to primaries. However, new words list without show the trend of words decreasing. the words should be recognized and those can be write, both of them should be commonly used words, the coverage of secondary used words is not very high , some words even appeared in textbooks , though have high frequency of use . Therefore teaching materials need to amendment, and increase some commonly used words. In addition, reduce amount of non-commonly used words in textbooks, delete some non-commonly used words according to actual teaching, it also can reduce the study burden to students.

5. Word'S Repetition And Reappearance Rates Of Words

Word's repetition refers to the frequency of the appearance of a word in the entire word corpus. The more times a word appears, the higher the rate of the reappearance is, and the better a student learns. The reappearance rate is the percentage of the word's repetition as a part of the total of all the Tibetan

words. Word's repetition and the reappearance rate of Tibetan words in primary school textbooks are shown in table 3.

Table 3 The distribution table of Tibetan words in primary school language textbooks

Numble	Tibetan	Repetition	Reappearance rate	Accumulated frequency
1	ལ	2642	2.45	2.45
2	ད	2562	2.37	4.82
3	ནས	2123	1.96	6.78
4	ཀྱི	1770	1.64	8.42
5	དང	1766	1.63	10.05
6	བ	1531	1.42	11.47
7	ཀུ	1310	1.21	12.68
8	ཚ	1258	1.16	13.84
9	ཐའི	1213	1.12	14.96
10	ཞེས	1082	1.1	16.06

In the textbooks, there are totally 2922 Tibetan words and the times of appearance of each word are extremely uneven. 598 words which only appear one time account for one fifth of the total Tibetan words, 637 words appear 2—4 times, 839 words appear 5—20 times, 634 words appear 20—50 times, 634 words appear 50—100 times, 182 words appear 100—500 times, and 32 words appear over 500 times. The uncommonly used words with high reappearance rate should have their reappearance rate reduced and the commonly used words with low reappearance rate should have their reappearance rate enhanced. Therefore enhancing the reappearance rate of the frequently used and easily-misused words can reduce the rate of error of the students.

6. Conclusion

According to the above analysis, Tibetan language textbooks of primary school collect 2922 words many of which are uncommonly used words and on the difficult side, each of Tibetan words with extremely uneven repetition. Therefore, in addition to taking account of the coverage of commonly used words, the new words learning of Tibetan language in primary schools should needs further in-depth study of problems from all aspects of teaching and learning, especially the cognitive laws of students' learning of new words. Compile Tibetan language textbooks with literacy teaching as the key link for lower and middle grades and make teaching materials of primary schools be able to meet the needs of the development of Tibetan basic education and even the entire Tibetan education.

Acknowledgements

This aritcle is subsidized by The National Natural Science Foundation of China (Grant No. 61262052)

References

[1] Zhou Wei: On the Tibetan language policy and the development of Tibetan language education, Journal of Tibet University, Vol. 22 (2007) No.4, p.75-81.

[2] Zhang Xiuqin, Liu Jun, Zhu Shaohui, et al. A research report of the currently used textbooks of the Tibetan language in Qinghai province, Ethnic Education Study, Vol. 15 (2004) No.6, p.36-43.

[3] Su Hong: Discussion on compiling Chinese language teaching material in primary and secondary schools for Tibetans , Qinghai Education, Vol. 12(2004) No.12, p.20-21.

[4] The Tibetan language teaching materials leading collaboration group: *Tibetan Language In Primary School* (Xining: Qinghai Nationalities Publishing House, China 1982).