# The Metering Research of Words in 2011 Tibetan Newspapers

H. Cao, J.S. Wei, Z.X. Leng Ben, W. Yang
Key lab of China's National Linguistic Information Technology
Northwest University for Nationalities
Lanzhou, China

***Abstract*—With the rapid development of social life, Tibetan language life is also constantly changes, a variety of contradictions in language are obvious. To grasp the dynamic language comprehensively and accurately could be better to research and use this kind of language resources, also could achieve the goal of protecting the language ecology and creating a harmonious language life. This article focuses on words of the Tibetan newspapers, survey the situation of the words about coverage, Frequency, Frequency segmentation statistic, high-frequency words and new words. This study not only understands the using conditions of these words and hot topics in Tibetan areas and ethnic customs, but also reflects the Tibetan language life to a certain extent.**

***Keywords-language life; Tibetan newspapers; using condition***

## I. INTRODUCTION

In today's modern life, mess media is increasingly influencing people's daily language and various aspects of social life. Tibetan newspapers have one hundred years of development history; have become influential in public communication system in China. They are popular in Tibetan areas.

Language is the most important information carrier and communication tools [1].Research on words of Tibetan newspapers could timely reflect the Tibetan language conditions, describe the Tibetan language change and provide people with the actual situation of Tibetan language. Research of words in Tibetan newspapers could help people understand Tibetan social language life through public opinion, provide people with the actual situation of Tibetan language, life hotspot and difficulty .It also can achieve the goal of protecting the ecology，create a harmonious language life and realize the sustainable development of language resources[2].

## II. CONTENT AND METHOD

The selected corpora used in this survey (sort by the founded time of the Tibetan newspapers): <Qinghai Tibetan News>, <Gannan Tibetan News>, <Tibet Daily> and <Tibet Science and Technology Newspaper>.

The collection of the corpus combined with labors and computer software. The word segmentation uses the method of combining machine-automatic segmentation and manual intervention. In words segmentation, we used a kind of based on statistics "Tibetan segmentation software" written by members of the group.

## III. THE SITUATION OF USING WORDS

In this research, the number of text corpus is 14784, the number of <Qinghai Tibetan News> is 4639, the number of <Gannan Tibetan News> is 2658, the number of <Tibet Daily > is 4883, the number of <Tibet Science and Technology Newspaper> is 2604.The species number of words is 19119, and the number of words is 5241895.

### A. *Coverage*

Coverage is one of the important indicators and reflects the usage of words. Coverage is the specified survey within the corpus as a percentage of the total amount of survey [3]. Figure 1 shows the word distribution of the coverage of 10%-90% and 91 %-100% of each segment.

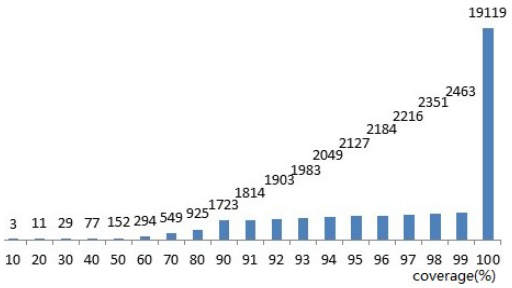


FIGURE I. THE DISTRIBUTION OF THE NUMBER OF WORDS IN A DIFFERENT COVERAGE

It can be seen, the highest frequency of the first three words' coverage up to 10%，illustrates the usage rate of the three words in the newspaper is quite high. 925 words covered 80% of corpus, 1723 words covered 90%，the words whose coverage reached 90% are called high-frequency words. 2463 words who accounted for 12.88% of the total words covered 99% of corpus. There are many words whose coverage rate were between 99% and 100%.

### B. *Frequency*

Frequency is the number of occurrences of the survey in all corpora. Frequency can directly reflect the words' use conditions. Table 1 shows the number of corresponding words in different frequency.

TABLE I. THE NUMBER OF WORDS IN DIFFERENT FREQUENCY

| Frequency | Number of words | The proportion（%） | Accumulative（%） |
|---|---|---|---|
| **1** | 2761 | 14.44 | 14.44 |
| **2** | 1394 | 7.29 | 21.73 |
| **3** | 1013 | 5.30 | 27.03 |
| **4** | 854 | 4.47 | 31.50 |
| **5-10** | 2887 | 15.10 | 46.60 |
| **11-20** | 2178 | 11.39 | 57.99 |
| **21-100** | 4087 | 21.38 | 79.37 |
| **>100** | 3945 | 20.63 | 100.00 |

The proportion of the Tibetan high-frequency words is low; the number of low frequency words is big. The frequency of no more than four accounted for 31.5% of the total number of words, the proportion of words whose frequency more than 100 is 20.63%, these show the proportion of high-frequency words is low, low frequency words quantity is big.

### C. *Frequency Segmentation Statistic*

According to the word frequency from high to low sorting of words, divided it into four frequency sequences. Table 2 shows distribution of the newspapers in word frequency sequences.

TABLE II. THE DISTRIBUTION OF WORDS IN FREQUENCY SEQUENCE SEGMENT

| Frequency sequence segment | The ratio of Words（%） | Frequency | The coverage of frequency（%） |
|---|---|---|---|
| **1-1000** | 5.23 | 4263655 | 81.34 |
| **1001-3000** | 10.46 | 655887 | 12.51 |
| **3001-10000** | 36.61 | 300300 | 5.73 |
| **>10000** | 47.70 | 22053 | 0.42 |

It can be seen that although the newspaper top 1000 words only account for 5.23% of the proportion of the word species, their coverage has reached 81.34%, the rest of the words although account for 94.77% of the proportion of the word species, their coverage is only 18.66%.

### D. *High-Frequency Words*

High-frequency words are words whose coverage rate reaches 90%. There are 1723 high-frequency words in the Tibetan newspapers (see table 1).

*1) The length of the word:* Tibetan language composed of syllables, the length of the word is the word's syllables number. The length of the high-frequency words in newspapers is as shown in figure below.

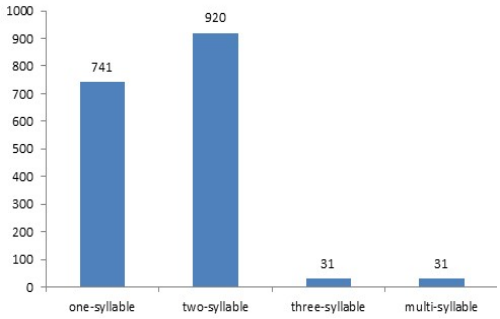


FIGURE II. THE DISTRIBUTION OF THE LENGTH OF THE WORD

It can be seen from the Figure 2,two-syllables words account for the largest proportion, one-syllable words followed by three-syllable words and multi-syllables words. account for the smallest proportion. Both the proportion of three-syllable high-frequency words and multi-syllables high-frequency words are low, this reflects the newspapers as mass media, in order to improve the popularity and readability, often use relatively brief words.

*2) Share and use alone:* High-frequency Tibetan words shared are the words commonly used in the four newspapers within their respective high-frequency words; the words that used alone of the high-frequency is referring to just using in one newspaper. The shared and used alone condition of the high-frequency words in newspapers are as shown in figure below.

TABLE III. THE WORDS OF FOUR NEWSPAPERS SHARED AND USED ALONE

| Media | The number of words | The number of the high frequency words | The number of shared words | Share Proportion（%） | The words used alone | Use alone proportion（%） |
|---|---|---|---|---|---|---|
| **1** | 12 509 | 2 361 | 1 021 | 43.24 | 303 | 12.83 |
| **2** | 14 552 | 2 546 | | 40.10 | 422 | 16.58 |
| **3** | 9 595 | 2 013 | | 50.72 | 434 | 21.56 |
| **4** | 13 968 | 1 639 | | 62.29 | 109 | 6.65 |

Gannan Tibetan News is Media 1, Qinghai Tibetan News is 2, Tibet Science And Technology Newspaper is 3, Tibet Daily is 4. It can be seen from the Table 3, the four newspapers' share high-frequency count is 1021, and sharing ratio is between 43.24% and 62.29%. The biggest amount of the words used alone is 434, the minimum is 109. <Tibet Science and Technology Newspaper> has the highest use alone proportion (21.56%), <Tibet Daily> has the lowest use alone proportion (6.65%). The audiences of <Tibet Daily> are the most wide, Selected the high-frequency words relatively is fixed and concentrated, but <Tibet Science and Technology Newspaper> as a newspaper focuses on science and technology, Spreads propaganda policy of science and technology, advances production technology, promotes new achievements of science and technology, therefore it uses the words to reflect these characteristics.

*3) New words:* New words of Tibetan adopt the following new words in the dictionary as a reference. If one

word appears in the following dictionary, we will take it as a new word. The first source is the new terminology dictionary, which contains more than 5300 new Tibetan words, made up in November 2009 edited by the Tibetan Chamber of Chinese National Language Translation Centre, and 587 new words has been explicated and added in 2012. The second resource is new words commonly used graphic dictionary in the Chinese-Tibetan-English which edited by Khenpo Cicheng Luozhu, at the same time we reference to new words commonly used dictionary edited by Khenpo Cicheng Luozhu. The third source is more than 550 new words issued by Tibetan Terminology Standardization Authorized Committee of Qinghai province from the first issue to the twentieth issue. The third resource is 2145 new words has been explicated by Professor Hua Kan of the Northwest University for Nationalities.

New words of newspapers can be roughly divided into seven areas. There are many new words in the field of economy such as དབང་གཉེར་ལེ་ལས་ཡ། (private entrepreneur) དཔལ་འབྱོར་མཁས་རྟན (economist) མ་དངུལ་འགེར་རྒྱག (turnover of capital) བར་སྣེན་ཡ། (economic man) ཤང་དབང་། (stock right) མ་ཤང་ནང་ ཞུགས་ཡ། (equity participation) བསྒྱར་བཅོས་སྒོ་དྲྭ། (reform and open) ཇོ་འགོད་ མ་རུ། (registered capital) འཐབ་ཏུས་ཀྱི་གོ་སྐབས་གལ་ཆེན། (important period of strategic opportunities) དཔལ་འབྱོར་ཚོང་མཉམ་ལས། (economy and trade cooperation) དཔལ་འབྱོར་གྱི་གྲལ་ཆ། (economic structure)and so on. There are many new words in the field of science and technology such as གློག་རྩལ་རྩེས་འགེར། (electronic computer) ཐོག་ཚུང་བྱ་རྒྱ། (local area network) སྣེན་མང་གཟུགས། (multimedia) སྐྱེ་དངོས་ལག་རུལ། (biotechnology) ཚན་རིག་འདུས་ཚོད། (technology content)and so on. There are many new words in the field of farming and animal husbandry such as འབྱོག་ཇོག (animal husbandry products) སྐྱེ་ཕུན་སྒྱག་ལས། (organic animal husbandry) ཞིང་ ལས་ལག་རུལ་ཆེ་སྐྱ། (agro-technician) འདེབས་འདྲུག་ལས་རིག་གྱི་གྲལ་ཆ། (planting structure)and so on. There are many new words in the field of tourism such as གནས་སྐོར་ལེ་ལས། (tourism enterprises) གནས་སྐོར་ལས་གཞི། (tourism) ནས་སྡེད་ས། (view spot) འགྲུལ་སྐྱོད། (travel) ཡུལ་སྐོར་ཐོན་ཞུངས། (tourist resources)and so on. There are many new words in the field of legal system such as ཁྲིམས་སྐྱོང་ཁོར་ཡུག (legal environment) ཁྲིམས་ལུགས། (law)and so on. There are many new words in the field of culture such as རིག་གནས་སྣ་མང་རུན། (culture diversity) ས་ཁོངས་རིག་གནས། (regional culture)and so on. There are many new words in the field of sports such as ཨོ་ལིན་ཏུ་སྐ་ཁེ། (Olympic)and so on.

## IV. Conclusions

This paper has researched words of Tibetan newspapers including the coverage, frequency, high-frequency words, the length of the word and new words, shared word, the words used alone and new words, effectively reflects the Tibetan newspaper words usage. But there are still two disadvantages. Firstly, the source of the corpus is single, Corpus quantity is small, which causes the persuasiveness of research data is inadequate. Secondly, the history of language can be researched from two angles of synchronic and diachronic [4], but this article only researched the Tibetan words of newspapers form the synchronic point of view. In subsequent work, I will further solve the above problems, to ensure the research could better reflect the Tibetan words usage.

## Acknowledgment

This paper received the support of Northwest university for nationalities special funding for basic scientific research (31920140057 ) .

## References

[1] J. Ma. The penetration of network words to daily languages [J].Journal of northwestern polytechnic university (Social Science Edition), 2002:53-57.

[2] National language resource monitoring and research center. The report of Chinese' language life situation in 2006 (next series)[M]. Beijing: the commercial press, 2007.

[3] M. Hou. Language resources construction and life monitoring related terms introduction [J]. Journal of terminology standardization and information technology, 2010, (2) : 30 to 33.

[4] Y.M. Li. View of language resources and language census in China [J]. Journal of Zhengzhou university (philosophy and social science edition), 2008, (1) : 5-7.