INASS

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Attribute Analysis in Basic Components of Tibetan Word

Cai Zhi Jie[*]**, Cai Rang Zhuo Ma**

Tibetan Intellectual Information Processing Centre of Qinghai Normal University, Xining, Qinghai, China

[*] *Corresponding author's Email: czjqhsd@163.com*

Abstract: Tibet has a long and glorious cultural history, and the number of its abundant historical documents is only less than that of Han Race in China, of which Tibetan words were created during the 7th century and have witnessed large-scaled normalization three times in the history. As a statistical analysis of words' properties provides basic data for the study of words, the statistical analysis of Tibetan words' properties can be considered as a fundamental task of Tibetan information processing. This paper carries out the frequency statistics of basic components from 2,129,234 Tibetan corpuses, and analyses the statistical data. It is the first time to conduct an analysis on basic components of large modern Tibetan corpus, and it has important value for the research of Tibetan linguistics and Tibetan information processing.

Keywords: Tibetan information processing; Corpus; Tibetan word; Basic component; Frequency.

1. Introduction

The statistical analysis of languages has a long history, the first frequency dictionary in the world is *The Frequency Dictionary of German* compiled by F. W. Kaeding in 1898 [1]. In China, the statistical analysis on Chinese sprang up in about the 20th century, and in 1970s, institutes (including Beijing Xinhua Enterprise) compiled *The Frequency Table of Chinese Characters* [2] through counting Chinese characters' frequency on 2.1 millions characters. In 1983, the researchers of Beijing Institute of Aeronautics and Astronautics conducted a large-scaled statistical analysis on modern Chinese language via computer, and have gained 13 character frequency tables [3]. This research on the usage of characters has provided objective data for information processing. Hence, the frequency analysis of Tibetan characters will not only provide important data for the quantitative research of Tibetan word, but also have important guiding and referential value for Tibetan education and Tibetan information processing, which makes the property analysis of Tibetan word a fundamental work of Tibetan information processing [4]. Chinese Research Centre of Tibetan (CRCT) has done fundamental work in Tibetan character frequency analysis, and given the frequency statistics of *Gangyur of Chinese of Tripitaka* [5]; Tibet University has made a statistical study on the qualities of all modern Tibetan character set [6]; Northwest University for Nationalities has given the statistics of frequency of the characters and syllables, and the entropy of Tibetan characters, syllables and the absolute entropy [8][9]. However, all the above works have not deeply analyzed the components of Tibetan words, so based on the work of the component decomposition of Tibetan characters [10][11], this paper describes a statistical model for the component frequency of modern Tibetan characters, and analyses the basic components of Tibetan characters from the corpus with 2,129,234 characters.

2. The Statistical Model for Tibetan Word Frequency

2.1 Model tibetan word

Tibetan language is the major communication tool in Tibetan regions, and it is gradually improved with the development of society. Modern Tibetan is a special alphabetic string that follows modern Tibetan writing grammar, and its main components are consonants. Modern Tibetan consists of the following parts: basic consonants, former letters, head letters, under letters, later letters, backmost letters and vowels. Statistically, modern Tibetan includes 30 basic consonants and 4 vowel letters, of which 10 consonants are later letter, 5 consonants are former letters, and 2 are backmost letters. Tibetan word's structure has both horizontal and vertical direction writing types—that is, former letters, root letters, later letters and backmost letters are horizontally written, and the root letters of vertical direction are head letters, under letters and vowels. The structure of Tibetan word is shown in figure 1.

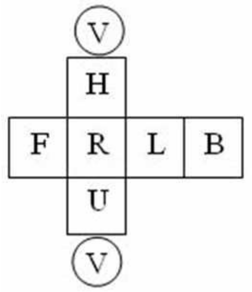


Figure1. The structure of Tibetan word.

In accordance with the grammar of modern Tibetan, Tibetan words have 1 to 7 components, and each part of the word is called component; each single former letter, head letter, root letter, under letter, vowel letter, later letter or backmost letter is called the basic component; in the vertical direction of the root letters, the whole character component composed of head letters, under letters and vowel letters is called the combined component (known as word butyl).

2.2 The statistical model for tibetan word frequency

When performing the frequency statistics on the components of Tibetan words, we firstly need to recognize and classify each symbol, such as English characters, numerical characters, graphic characters and others, and then decompose the components of Tibetan words, finally determine the positional features of each component in the word and its frequency statistics. The statistical model is shown in figure 2.

2.3 The library structure of word table

Word table library is the basic data of Tibetan word attribute analysis which includes basic component character table library, combined component character table library, coarse-grained character table library, fine-grained character table library and rule library.

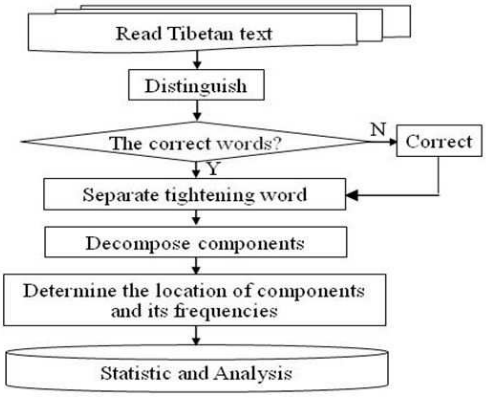


Figure 2. The statistical model for the components of Tibetan words.

BCDB (Basic Component DB) contains the basic components of modern Tibetan , including 30 consonants and 4 vowels, 3 head letters, 4 under letters, 10 later letters, 2 backmost letters ,5 former letters, and none-modern Tibetan characters which frequently appear in modern Tibetan. Its structure is shown as follows:

```
Typedef Struct BCDB {
  // Stores basic components.
  string [] ZW;

  // Describes the type
  // of basic components.
  int TYPE;

  // Describes the frequency
  // of basic components.
  int N;
};
```

ZW stores the basic components and non-Tibetan characters.

Modern Tibetan words contain 61 basic components. The frequency of the components stores in N. The value of TYPE is between 1-61.When the value is 1-30, which represents the component of ZW containing 30 consonants. When the value is 31-34, it represents the component of ZW contains 4 vowels. When the value is 35-37, it represents the component of ZW contains 3 head letters, 38-41 4 under letters respectively, 42-46 5 later letters, 47-56 10 later letters, 57-58 2 backmost letters, and 59-n (n $\geq$ R59, N can get different value according to the number of non-modern Tibetan word. In this system N=61) non-modern Tibetan words respectively. To improve the function of the system, this data base should be stored with partial order, in other words, the value of TYPE in certain range should follow the order of ZW. If we take a modern Tibetan word as a unity, then the word may have at most four components, and it has one and only one basic character or combined components.

According to the statistical and analysis of large Tibetan corpus, the total number of combined components in the modern Tibetan words are 476, the types and quantity of components as follows: vowel + root letter (120); head letter +root letter (33);root letter + under letter (36); vowel +head letter+ root letter (132); head letter + root letter + under letter (15); vowel +root letter+ under letter (95), vowel +head letter + root letter+ under letter (45).Because the quantity of combined components are countable and each type of component has smaller size, we can establish an orderly word table library of combined components-CCDB (Combined Component DB), and decompose combined components into basic components. The structure of combined component DB (CCDB, for short) is defined as:

```
Typedef Struct CCDB {
  // Stores combination components.
  string[] FHGJ;

  // Describes the root letters
  // of components.
  int TAG1;

  // Describes the vowels
  // of components.
  int TAG2;

  // Describes the head letters
  // of components.
  int TAG3;

  // Describes the under letters
  // of combined components.
  int TAG4;

  // Describes the most-under letter
  // of combined components.
  int TAG5;

  // Describes the frequency of
  // combined components.
  int N;
};
```

CCDB is composed of combined components and the flags of root letters, vowels head letters, under letters and most-under letters. It is used to store the combined components and the flag message of each part of combined components.

FHGJ field holds the combined components, TAG1 ($0 \leq$ TAGI ≤ 30) stores the natural sequence of 30 basic letters, TAG2 ($0 \leq$ TAGI ≤ 4) indicates if the combined component has vowels and which one,TAG3 ($0 \leq$ TAGI ≤ 3) describes if the combined component has head letters or not , TAG4 ($0 \leq$ TAGI ≤ 4) describes if the combined components has under letters or point it out,TAG5 ($0 \leq$ TAGI ≤ 4) indicates if the combined components has most-under letters or recognize it. Different I in TAGI are used to represent root letters, vowels, head letters and under letters.

The basic structure of Tibetan words is [former letter +] root letter / combined component[+ back letter][+ most-back letter], and it must contain the root letter or combined component. The structure of Tibetan words can be analyzed in two ways: if take the combined components as a unit, there are 11 coarse-grained structures. If decomposes further, there are 48 fine-grained structures. In order to analyze the attribute of Tibetan words from different perspective, we establish Coarse-grained Structure Character Property Database (DSDB for short) and Fine-grained Structure Character Property Database (FSDB for short). The DSDB and FSDB can be in disorder, their structures are defined as:

```
Typedef Struct DSDB {
  // Stores coarse-grained structure.
  string[] STRUCTURE;

  // Describes the frequency
  // of the structure.
  int N;
};

Typedef Struct FSDB {
  // Stores fine-grained structure.
  string[] STRUCTURE;
```

```
    // Describes the frequency
    // of the structure.
    int N;
};
```

The rule library is used to correct Tibetan words and bound the analysis of the Tibetan word attributes, which contains the rule of word constituting of modern Tibetan words, the decomposing regulation of different attribute and the descriptive rule of word structure.

In order to ensure the validity and common use of this system, the rule database uses an open structure, and the user can edit the rule while using the system.

3. The Frequency of the Basic Component of Tibetan Words

3.1 The frequency statistics of the components of tibetan words

Modern Tibetan is composed of basic components in the left-and-right and the top-and-down manners. The basic components include 30 consonants characters, 4 vowel characters, 3 head letters, 4 under letters, 5 former letters, 10 later letters and 2 backmost letters. This paper provides the frequency of the basic components from a corpus containing 2,129,234 Tibetan characters that covers *Selected Works of Deng Xiaoping* (Volume II), *Selected Works of Jiang Zemin* (Volume III) ,*The Constitutional Law*, law files and net resources. The frequency statistics of the basic components in the corpus is shown in table 1 to 8, the frequencies of the root letters, vowels, head letters, under letters, former letter, later letter and backmost letter are shown in figure 3 to 10 respectively.

Specifically, table 1 only shows the statistics result of consonants frequency and does not differentiate the different position message of the same consonant; Table 2 only shows the frequency of the basic letters of a consonant after removing the former letters, back letters and most-back letters.

3.2 Statistical analysis

Table 1-8 indicate that, except the root letters and backmost letter other components all are widely used in the corpus. The frequency statistical result of the basic components is as follows: For root letters, the most common are, which accounts for about 11.53%, 9.41%, 8.23% and 7.57% respectively; component and only accounts for nearly 0% and 0.03% respectively, which take small proportions. According to the proportions they take, other root letters can be divided into 4 groups, of which the frequency is consistent generally inside each group.

The first group contains the root letters ག་ད་བ་པ(whose proportion is larger than 7); The second group contains the root letters ཀ་ལ་འ་ཁ་ཏ་མ་ན་ས་ཚ་ར(whose proportion is in between 3.13-4.99); The third group contains the root letters ཐཉཡཆཅཕངཙཤཛཞ(whose proportion is in between 1.13-2.9); The fourth group contains the root letters ཝ་ཟ་ཇ་ཧ་ཨ (whose proportion is in between 0-0.9).

In the 2,129,234 characters Tibetan corpus, vowels appear 675151 times, where accounts for 31.82%, accounts for 21.55%, accounts for 16.39%, accounts for 30.24%. Head letter appear 148324 times, where accounts for 40.27%, accounts for 8.91%, ས accounts for 50.82%: the frequency is not very consistent, ལ occurs rarely. Under letters appear 236526 times, where ྱaccounts for 69.38%,ྲ accounts for 25.54%,ླ accounts for 4.48%,ྭ accounts for 0.60%: the frequency is not consistent. Former letters appear 249330 times, where the frequency is basically consistent. Later letters appear 766420 times, whereང accounts for 21.25%, ས accounts for 14.89%,ན accounts for 13.80%,ག accounts for 13.56%, which is a larger proportion,འ accounts for 0.80%,མ accounts for 4.64%,བ accounts for 4.94%, which is a small proportion.

From the analysis above, first, the frequencies of the basic components of the Tibetan characters are not totally the same. Some basic components have consistent frequency, while others not; second, from the frequency of the components, the keyboard layout designed currently is not very reasonable. For example, the frequency of ཇ་ཧ་ཨ is very low, but the currently-designed keyboard layout puts them into district 3, while པ་ཀ་ཁ occur frequently, but they are put into district 4.

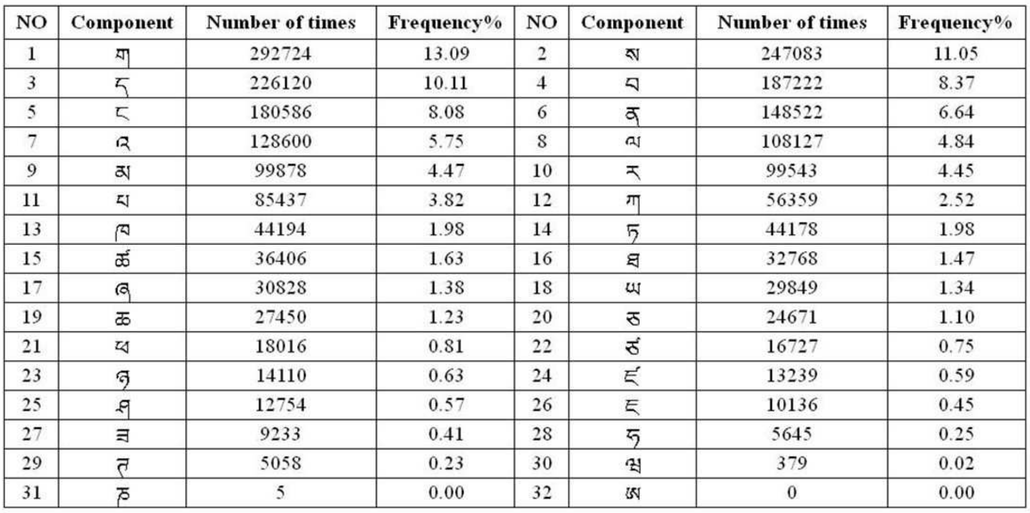
4. Conclusion

We conducted thorough statistical analysis on the basic components of Tibetan words which provides statistics for the Tibetan information processing. In the future, we would like to perform statistical analysis based on this research, so as to provide the basis for the research of the information entropy of Tibetan characters, and recognize the Tibetan phrase according to an N-gram language model.

References

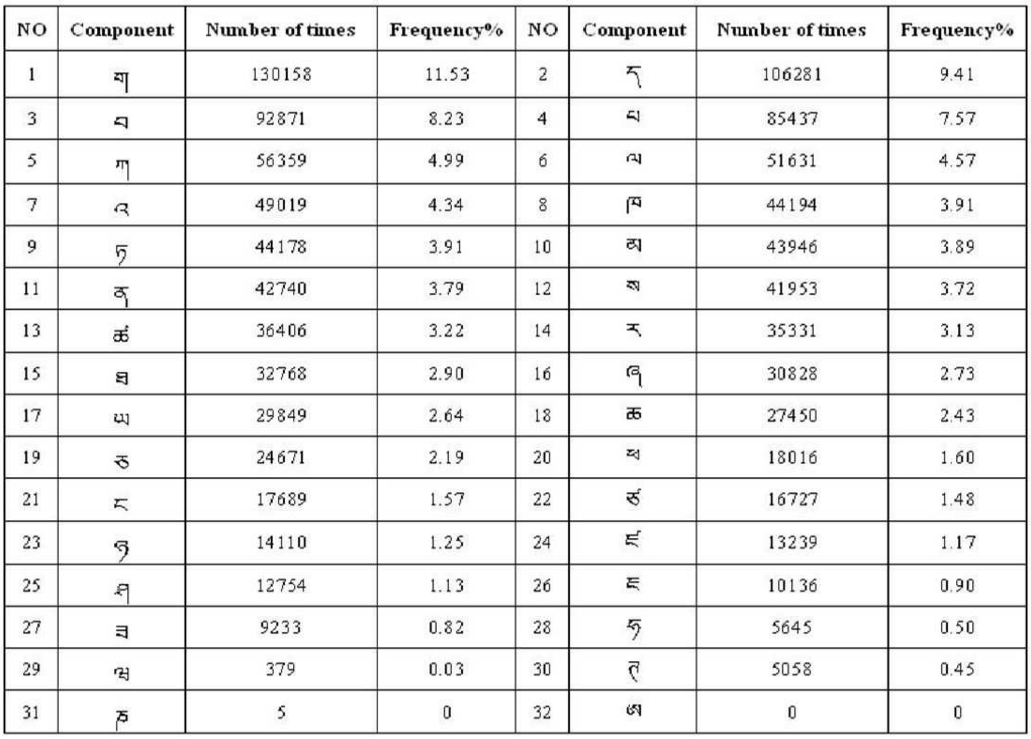
[1] Wang Guoquan, *Study on Virtual Test of Vehicle Ride Comfort, China Agricultural University*, Doctor of Science thesis, 2002.

Table 1. Statistics of consonants character frequency



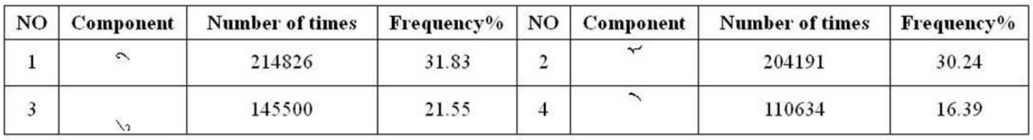
NO	Component	Number of times	Frequency%	NO	Component	Number of times	Frequency%
1	ག	292724	13.09	2	ས	247083	11.05
3	ད	226120	10.11	4	བ	187222	8.37
5	ང	180586	8.08	6	ན	148522	6.64
7	འ	128600	5.75	8	ལ	108127	4.84
9	མ	99878	4.47	10	ར	99543	4.45
11	པ	85437	3.82	12	ཀ	56359	2.52
13	ཁ	44194	1.98	14	ཏ	44178	1.98
15	ཚ	36406	1.63	16	ཐ	32768	1.47
17	ཞ	30828	1.38	18	ཡ	29849	1.34
19	ཆ	27450	1.23	20	ཅ	24671	1.10
21	ཕ	18016	0.81	22	ཙ	16727	0.75
23	ཉ	14110	0.63	24	ཛ	13239	0.59
25	ཤ	12754	0.57	26	ཇ	10136	0.45
27	ཟ	9233	0.41	28	ཧ	5645	0.25
29	ཊ	5058	0.23	30	ཥ	379	0.02
31	ཋ	5	0.00	32	ཨ	0	0.00

Table 2. Statistics of base character frequency



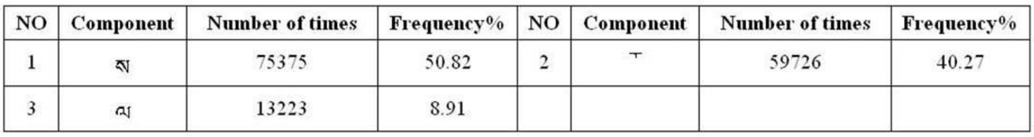
NO	Component	Number of times	Frequency%	NO	Component	Number of times	Frequency%
1	ག	130158	11.53	2	ད	106281	9.41
3	བ	92871	8.23	4	པ	85437	7.57
5	ཀ	56359	4.99	6	ལ	51631	4.57
7	འ	49019	4.34	8	ཁ	44194	3.91
9	ཏ	44178	3.91	10	མ	43946	3.89
11	ན	42740	3.79	12	ས	41953	3.72
13	ཚ	36406	3.22	14	ར	35331	3.13
15	ཐ	32768	2.90	16	ཞ	30828	2.73
17	ཡ	29849	2.64	18	ཆ	27450	2.43
19	ཅ	24671	2.19	20	ཕ	18016	1.60
21	ང	17689	1.57	22	ཙ	16727	1.48
23	ཉ	14110	1.25	24	ཛ	13239	1.17
25	ཤ	12754	1.13	26	ཇ	10136	0.90
27	ཟ	9233	0.82	28	ཧ	5645	0.50
29	ཥ	379	0.03	30	ཊ	5058	0.45
31	ཋ	5	0	32	ཨ	0	0

Table 3. Statistics of vowel character frequency table



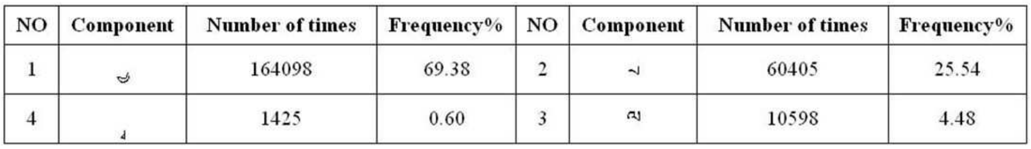
NO	Component	Number of times	Frequency%	NO	Component	Number of times	Frequency%
1	ི	214826	31.83	2	ོ	204191	30.24
3	ུ	145500	21.55	4	ེ	110634	16.39

Table 4. Statistics of head letter frequency table



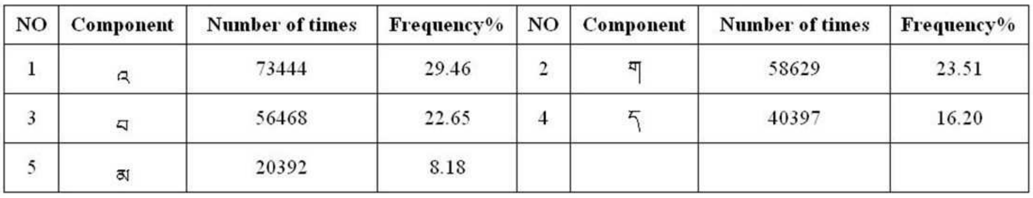
NO	Component	Number of times	Frequency%	NO	Component	Number of times	Frequency%
1	ས	75375	50.82	2	ར	59726	40.27
3	ལ	13223	8.91				

Table 5. Statistics of under letter frequency table



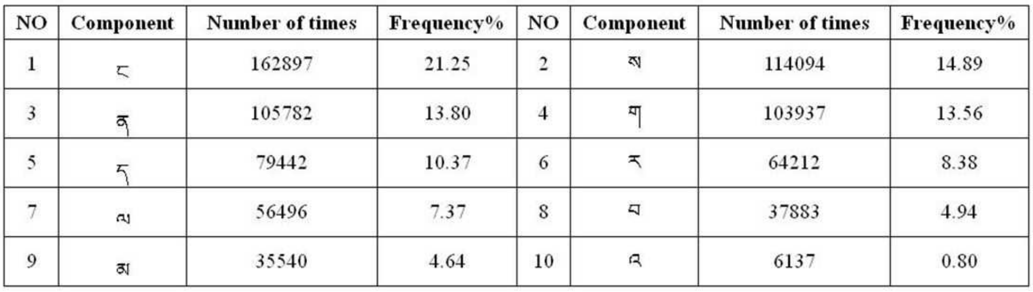
NO	Component	Number of times	Frequency%	NO	Component	Number of times	Frequency%
1	ྱ	164098	69.38	2	ྲ	60405	25.54
4	ྭ	1425	0.60	3	ལ	10598	4.48

Table 6. Statistics of former letter characters frequency table



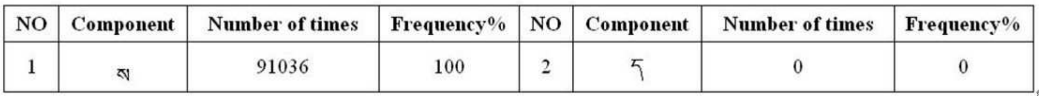
NO	Component	Number of times	Frequency%	NO	Component	Number of times	Frequency%
1	འ	73444	29.46	2	ག	58629	23.51
3	བ	56468	22.65	4	ད	40397	16.20
5	མ	20392	8.18				

Table 7. Statistics of later letter characters frequency table



NO	Component	Number of times	Frequency%	NO	Component	Number of times	Frequency%
1	ང	162897	21.25	2	ས	114094	14.89
3	ན	105782	13.80	4	ག	103937	13.56
5	ད	79442	10.37	6	ར	64212	8.38
7	ལ	56496	7.37	8	བ	37883	4.94
9	མ	35540	4.64	10	འ	6137	0.80

Table 8. Statistics of backmost letter characters frequency table



NO	Component	Number of times	Frequency%	NO	Component	Number of times	Frequency%
1	ས	91036	100	2	ད	0	0

[2] A. Kuznetsov, M. Mammadov, I. Sultan and E. Hajilarov, “Optimization of a quarter-car suspension model coupled with the driver biomechanical effects", *Journal of Sound and Vibration*, vol.330(12), pp.2937-2946, 2011.

[3] P. S. Els, N. J. Theron, P. E. Uys and M. J. Thoresson, “The ride comfort vs handling compromise for off-road vehicles", *Journal of Terramechanics*, vol.44(4), pp.303-317, 2007.

[4] X. Q. Zhang and B. Yang, “Simulation and Analysis of Ride Comfort under Random Road based on ADAMS", *Shanghai Auto*, (8), pp.18-23, 2011.

[5] Y. Zhang and A. Tang, “The CAE Revolution and the Development of the Virtual Proving Ground Approach", *The 4th International LS-DYNA Conference*, Minneapolis, Minnesota, 1996.

[6] D. C. Lee and C. S. Han, “CAE (computer aided engineering) driven durability model verification for the automotive structure development", *Finite Elements in Analysis and Design*, vol.45(5), pp.324-332, 2009.

[7] C. R. Lee, J. W. Kim and J. O. Hallquist, “Validation of a FEA Tire Model for Vehicle Dynamic Analysis and Full Vehicle Real Time Proving Ground Simulations", SAE971100

[8] G. Q. Guang and Y. Fang, “Time-Domain Simulation and Analysis of Vehicle Ride Comfort", *Automobile Technology*, vol.8(2), pp.8-11, 2007.

[9] Changchun Automobile Research Institute, “GB/T4970-1996 Method of random input running test-Automotive ride comfort", 1996.

[10] H. Zhao and S. F. Lu, “Road input in time domain model on four tires vehicle", *Automotive Engineering*, vol.21(2), pp.112-117, 1999.

[11] J. Yang, Y. Suematsu and Z. Kang, “Two-degree- of-freedom controller to reduce the vibration of vehicle engine-body system", *IEEE Transactions on Control Technology*, vol.9(2), pp.295-317, 2001.

[12] R. A. Williams, “Automotive active suspensions", *Part 1: basic principles*, *Proceeding of Institute of Mechanical Engineers*, pp.415-426, 1997.

[13] Q. Zhu and M. C. Ishitobi, “Chaos and bifurcations in a nonlinear vehicle model", *Journal of Sound and Vibration*, vol.275(5), pp.1136-1146, 2004.

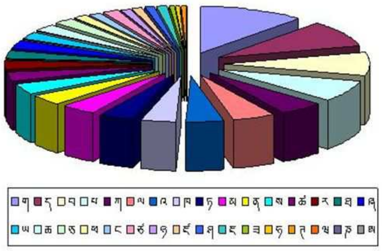


Figure 3. Frequency map of consonant letters

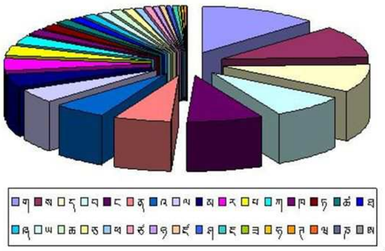


Figure 4. Frequency map of base letters

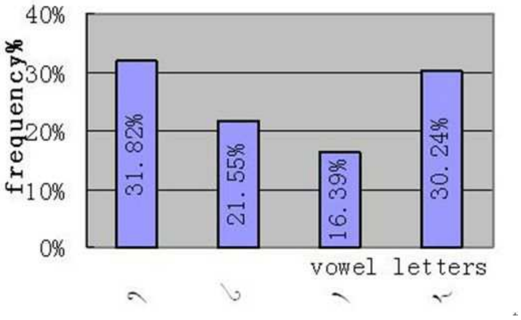


Figure 5. Frequency map of vowel letters

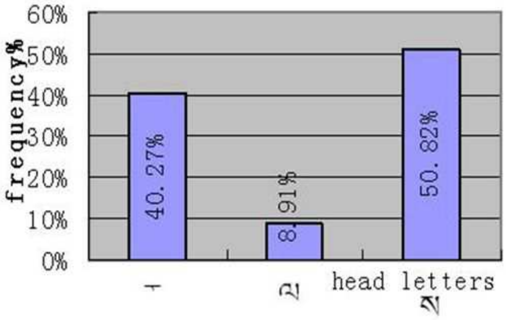


Figure 6. Frequency map of former letters

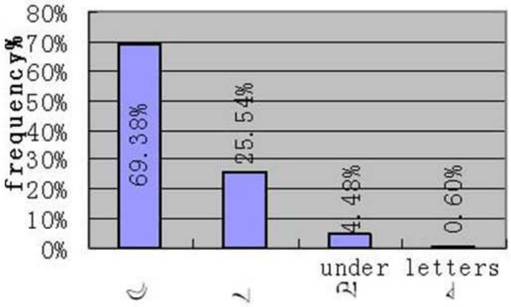


Figure 7. Frequency map of head letters

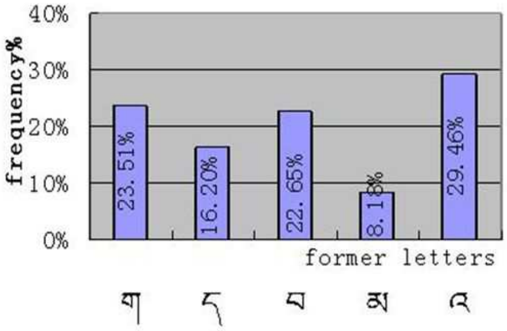


Figure 8. Frequency map of later letters

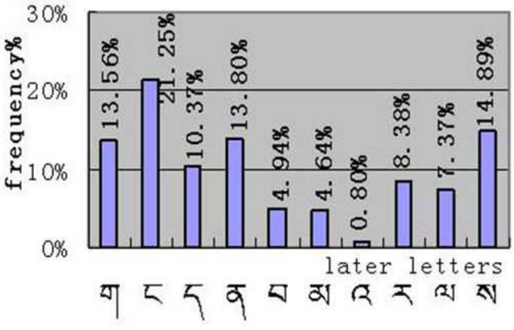


Figure 9. Frequency map of under letters

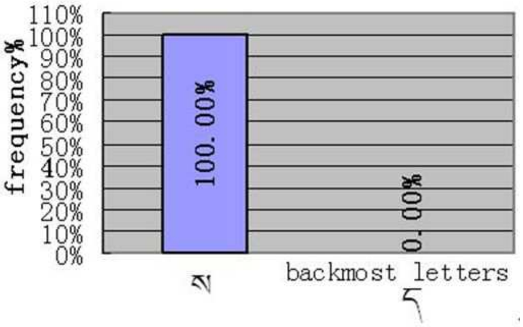


Figure 10. Frequency map of backmost letters

[14] M. Yamashita, K. Fujimofi, K. Hayakawa, et a1, "Application of H control to active suspension systems", *Automatica*, vol.30(11), pp.1717-1729, 1994.

[15] G. H. Tian, Y. S. Wang, A. L. Geng, "Time-domain linear simulation of automobile ride performance and software development", *Machinery Design and Manufacture*, vol.38(5), pp.80-82, 2006.

[16] W. Schiehlen and B. Hu, "Spectral simulation and shock absorber identification", *International Journal of Non-Linear Mechanics*, vol.38(2), pp.161-171, 2003.

[17] S. X. Gao, "Comment of Motor Vehicle Ride Vibration Evaluation", *Automobile Technology*, vol.12(3), pp.16-17, 1995.