International Journal of Knowledge and Language Processing
Volume 5, Number 1, **2014**

www.ijklp.org
KLP International ©2014 ISSN 2191-2734
pp.59—67

An Situation Investigation of Chinese Medicine Science Terminology Using in the Academic Journals

Liu Pengyuan[1] and Su Qi[2]
[1]Applied Linguistic Research Institute
Beijing Language and Culture University
liupengyuan@pku.edu.cn
[2]School of Foreign Languages
Peking University
sukia@pku.edu.cn

Received January 2014; revised January 2014

ABSTRACT. This paper introduces a statistical investigation to the Chinese medicine science terminology using in the academic journals. This is the first time as we know to do such terminology situation investigation research in Chinese science and technology terminology within medicine science terminology domain. It uses a Chinese medicine science terminology investigation list which including standard terminology and non-standard terminology. It mainly studies the general use situation in the terminology frequency of standard terminology and non-standard terminology. It made the comparison between standard terminology and non-standard terminology by statistical analysis. The data sources are 16978 papers which selected from China National Knowledge Infrastructure (CNKI). The investigation result shows that in Chinese medicine science academic papers: 1) The use of standard terminology has competitive advantage comparing to that of non-standard terminology (75%:25%) and 2) to express the same concept, 27% non-standard terminology have surpassed their corresponding standard forms.

Keywords: situation investigation, medicine science terminology, academic journals, standard terminology, non-standard terminology

1. **Introduction.** The science and technology terminology is mainly the science concept described by language. It is a kind of expression of language in the science and technology domain. The science concept needs precision and strictness more than the common concept every day we used, otherwise the information it disseminating and interchanging would be distortion and even become falsehood. Therefore science and technology terminology needs be standardized more. It would be a great obstacles to the development of science if science and technology terminology was used non-standardly and furthermore, it would

cause a great loss to our world because of the misunderstanding of science terminology.

It is very important in standardization and normalization of science terminology. Many researchers [1] studied it and they did a lot of outstanding work in this area. A survey of this research area in Chinese before year 2002 is presented in Feng Zhiwei [2]. It shows the history of the study of Chinese science terminology. Liu Qing [3, 4] studied the fundamental normalization process of Chinese science terminology and showed the important purpose in information dissemination procedure. Zhang Hui [5] interviewed Zhou Youguang and indicate that the normalization of science terminology is a key problem to the science development. Ma Juhong [6] did some discussion of the normalization of science terminology from a sociology point of view.

These researches focus on the necessity of the normalization of science terminology and make a lot of discussion on the method and the policy which can help the normalization work. It is very helpful to guide the work of normalization of science terminology. But these works basically are case study and lack statistical data to support especially they lack the macroscopic level use of science terminology in practical. Therefore we introduce the strategy of situation investigation based on the data of the terminology in practical use. Although there was a study on the investigation of Chinese Idioms and Idiomatic Phrases [7], it is the first time to do such investigation in Chinese medicine science terminology. We hope it can let us know the panorama of the use of the Chinese medicine science terminology and support the work of normalization of science terminology.

2. **Data preparation for the investigation.**

2.1. **Target investigation domain and paper selection**.

This paper focus on the medicine science domain. The investigation target is Chinese academic journal. The medicine science is equal to the Medical science and technology domain according to the classification of China National Knowledge Infrastructure (CNKI) which is the largest Chinese academic journal database. There are 28 subfields within the domain and a few hundred journals. There are about 4 million papers published in all these journals since 1949.

A subset of these papers as investigation sample were used in the paper. The gynaecology and obstetrics, stomatology, medical genetics, cardiovasology, ophthalmology, general surgery and plastic surgery these 7 subfields were chosen. We choose all 22 core journals of these 7 subfields as the sample source of the situation investigation.

We download 17463 papers[6] from these 22 journals from CNKI. The total size of the files in Portable Document Format is 6.41G bytes.

2.2. **Investigation Sample Dataset**.

2.2.1. **Preprocessing.** It is hard to manage because all of the files are in Portable Document Format (PDF files). We transfer all 17463 PDF files into TXT files. The total size of all the TXT files is 221.3M bytes. We delete all the short text files (It is generally call for paper

[6] All of these papers are selected during newest annual which included in CNKI as much as possible. We guarantee the integrity of each downloaded paper.

and notice to the reader) of them and get 16978 TXT files as our final investigation sample dataset.

Next we delete the redundancy characters (It is mainly blank) and the unreadable code which cannot be transfer from PDF pictures and tables and so on into the text files.

2.2.1. **The General Information of Investigation Sample Dataset.** The general information of investigation sample data-set is listed in table 1. The total amount of the Chinese characters are about 53 million of times.

TABLE 1. THE GENERAL INFORMATION OF INVESTIGATION SAMPLE DATA-SET

Total amount of Chinese characters	The amount of Chinese characters of longest paper	The amount of Chinese characters of longest paper	Average amount of Chinese characters per paper
52,959,218	53,198	501	3,119

A small terminology contrast list of standard terminology and non-standard terminology is established by a group medicine specialists. English standard terminology are regarded as concept in the terminology contrast list. There are standard terminology and non-standard terminology according to each English concept. In most cases, there is only one Chinese terminology contrast to one English concept but sometimes there are several Chinese terminology especially non-standard terminology are corresponding to one same English concept. The general information of standard terminology and non-standard terminology of computer science are listed in table 2. There are 258 English concepts corresponding to 266 standard terminology and 265 non-standard terminology.

TABLE 2. THE GENERAL INFORMATION OF TERMINOLOGY CONTRAST LIST

	Items number	Character amount of the longest item	Character amount of the shortest item	Average character amount
Standard terminology	266	14	2	4.14
Non-standard terminology	265	13	2	4.17

3. **Data Statistic and Analysis.** We perform statistical analysis to the investigation sample dataset based on the contrast terminology list. The results of frequency and distribution of terminology in the contrast terminology list are listed in table 3.

TABLE 3. GENERAL DISTRIBUTION INFORMATION OF STANDARD TERMINOLOGY AND NON-STANDARD TERMINOLOGY

	Total frequency	Maximum frequency	Minimum frequency	Average frequency	Ratio of Total
Standard terminology	226397	42049	1	13.33	75.0%
Non-standard terminology	75431	38611	0	4.44	25.0%

Table 3 shows that in general, the total frequency and average frequency of standard terminology both exceed that of non-standard terminology. The ratio is about equal to 3 (75% vs. 25% for standard terminology vs. non-standard terminology).

We compare the frequencies of standard terminology and non-standard terminology corresponding to the same concept in fig. 2. There are 183 concepts which the frequency of standard terminology are more than that of non-standard terminology. There are 71 concepts are just the opposite and 4 concepts is equal.

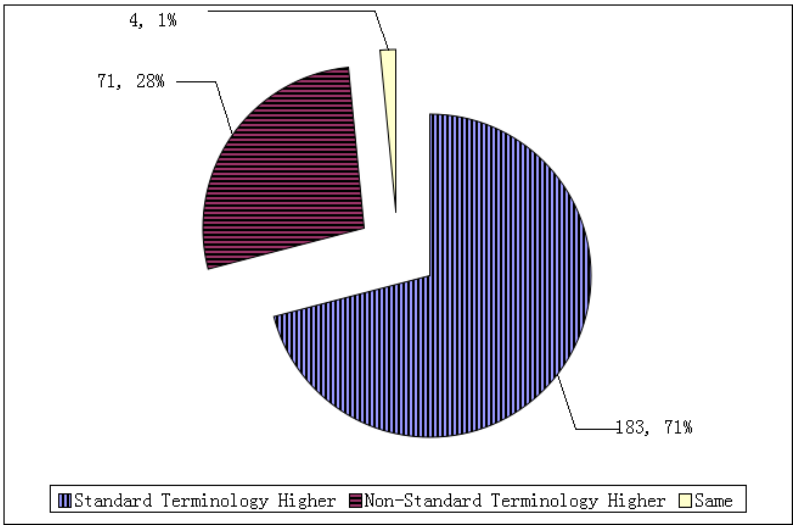

FIGURE 1. FREQUENCIES OF STANDARD/NON-STANDARD TERMINOLOGY CORRESPONDING TO THE SAME CONCEPT

In the same concept level, though there is a few differences from table 3, the similar pattern of frequency advantage of standard terminology is not changed (71% vs. 28% for standard terminology vs. non-standard terminology).

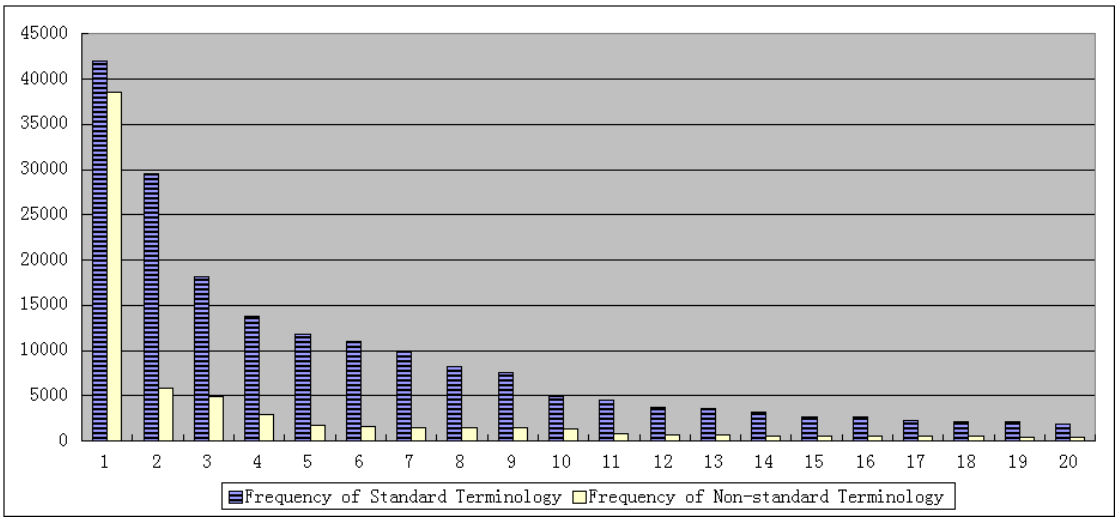

FIGURE 2. TOP 20 FREQUENCY OF STANDARD TERMINOLOGY AND THAT OF NON-STANDARD TERMINOLOGY

Arrange all the standard terminology frequencies and non-standard frequencies by descending order. Figure 2 is the bar chart of top 20 frequency of standard terminology and non-terminology. It shows that the frequencies of all most used standard terminology are more than that of non-standard terminology with the same rank.

TABLE 4. TOP 20 FREQUENCY OF STANDARD TERMINOLOGY AND THAT OF THEIR CORRESPONDING NON-STANDARD TERMINOLOGY WHICH SHARED SAME ENGLISH CONCEPT

Standard terminology	Frequency of standard terminology	Ratio of standard terminology	Non-standard terminology	Frequency of non-standard terminology	Ratio of non-standard terminology
妊娠	42049	98.66%	怀孕	569	1.34%
并发症	29554	95.24%	合并症	1477	4.76%
淋巴结	18206	99.98%	淋巴腺	4	0.02%
综合征	13838	99.13%	症候群	121	0.87%
胆管	11749	66.81%	胆道	5837	33.19%
晶状体	10941	69.01%	晶体	4913	30.99%
水肿	9831	98.64%	浮肿	136	1.36%
瘢痕	8242	96.93%	疤痕	261	3.07%
黏膜	7503	94.42%	粘膜	443	5.58%
死亡率	4856	62.64%	病死率	2896	37.36%
白细胞	4486	99.93%	白血球	3	0.07%
适应证	3762	99.58%	适应征	16	0.42%
红细胞	3557	99.89%	红血球	4	0.11%
胆总管	3154	99.43%	总胆管	18	0.57%
白蛋白	2667	63.64%	球蛋白	1524	36.36%
斜视	2650	99.66%	显斜	9	0.34%
胞浆	2304	80.36%	细胞浆	563	19.64%
血红蛋白	2135	97.40%	血色素	57	2.60%
胞质	2092	78.79%	细胞浆	563	21.21%
原发性高血压	1879	4.64%	高血压	38611	95.36%

The detail information of top 20 frequency of standard terminology and non-terminology which shared same terminology is listed in table 4. Figure 3 is the bar chart of the top frequencies comparison.

In Fig. 3 the Y-axis is frequency of standard/non-standard terminology and the X-axis is the rank of standard terminology frequency with descending order.

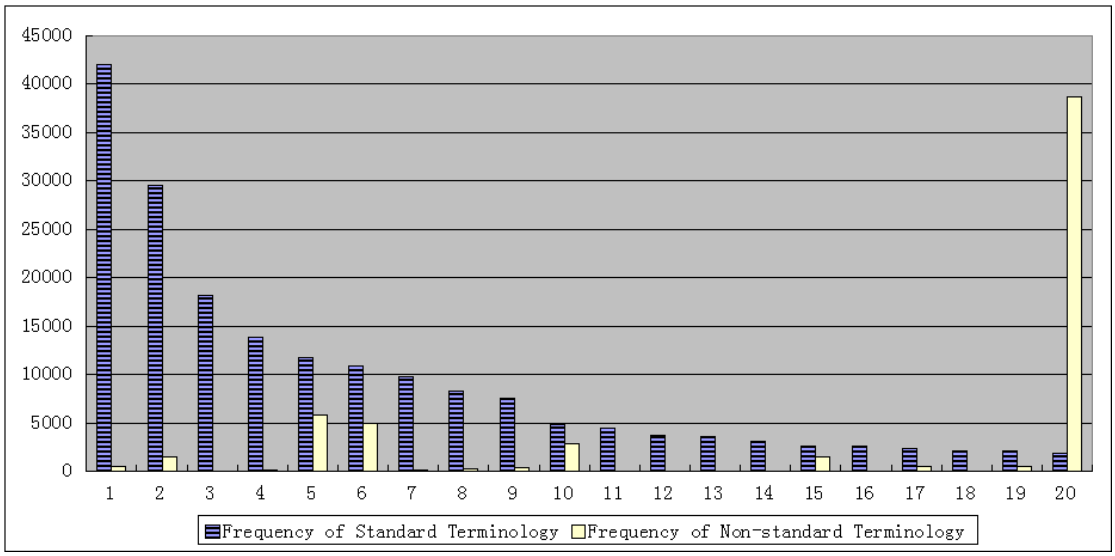

FIGURE 3. TOP 20 FREQUENCY OF STANDARD TERMINOLOGY AND THAT OF THEIR CORRESPONDING NON-STANDARD TERMINOLOGY WHICH SHARED SAME ENGLISH CONCEPT

Fig. 3 and table 4 show that all the frequencies of standard terminology are more than that of non-standard terminology which share the same English concept, except line 20th where the frequency of non-standard terminology is far more than that of standard terminology. The frequencies of 5 non-standard terminology (25%) are higher than 50% frequencies of standard terminology. It seems so abnormal that the frequency of line 20th which the standard terminology is "原发性高血压" (The English concept is "essential hypertension"). We can see that the corresponding non-standard terminology "高血压" is of a part of "原发性高血压" and is much shorter than "原发性高血压". At the same time the terminology "高血压" is a frequently used standard terminology of English concept "hypertensive/hypertension" Our investigation did nothing to distinguish different senses because the word sense disambiguation technique is not practical to most of natural language processing task [8] including this task. This drawback of our investigation lead to inaccuracy result but fortunately it is just only a few cases.

Afterwards we observe the frequency of non-standard/standard terminology along the rank of non-standard terminology frequency with descending order as showed in Fig.4 and table 5. From Table 4, 5 and Fig. 3, 4 we know that the distribution of standard terminology and non-standard terminology which shared same English concept is hard to form a suited rule. Table 5 and Fig. 4 show that there are about 11 standard terminology being used more frequent than non-standard terminology. At the same time we should realize that the rest other 9 non-standard terminology are used more frequent than standard terminology. Some standard terminologies such like "副流行性感冒" ("parainfluenza"),"代偿失调" ("decompensation"),"破骨细胞瘤" ("osteoclastoma") are very few. These standard terminologies are only used 1, 4, 1 times separately whereas each of these concepts is used over 500 times. It means that researchers prefer to use these non-standard terminologies in practice though they are not standard terminologies.

TABLE 5. TOP 20 FREQUENCY OF NON-STANDARD TERMINOLOGY AND THAT OF THEIR CORRESPONDING STANDARD TERMINOLOGY WHICH SHARED SAME ENGLISH CONCEPT

Non-standard terminology	Frequency of non-standard terminology	Ratio of non-standard terminology	Standard terminology	Frequency of standard terminology	Ratio of standard terminology
高血压	38611	95.36%	原发性高血压	1879	4.64%
胆道	5837	33.19%	胆管	11749	66.81%
晶体	4913	30.99%	晶状体	10941	69.01%
病死率	2896	37.36%	死亡率	4856	62.64%
磁共振	1662	72.36%	磁共振成像	635	27.64%
球蛋白	1524	34.35%	清蛋白; 白蛋白	2913	65.65%
全麻	1514	60.97%	全身麻醉	969	39.03%
合并症	1477	4.76%	并发症	29554	95.24%
介质	1468	76.98%	递质	439	23.02%
恶变	1348	95.54%	恶性变	63	4.46%
高脂血症	796	49.38%	脂血症	816	50.62%
分裂相	666	82.02%	分裂象	146	17.98%
流感	662	99.85%	副流行性感冒	1	0.15%
怀孕	569	1.34%	妊娠	42049	98.66%
细胞浆	563	9.32%	细胞质; 胞质; 胞浆	5477	90.68%
高血脂	560	40.70%	脂血症	816	59.30%
失代偿	553	99.28%	代偿失调	4	0.72%
骨巨细胞瘤	504	99.80%	破骨细胞瘤	1	0.20%
粘膜	443	5.58%	黏膜	7503	94.42%
外固定架	411	47.35%	外固定器	457	52.65%

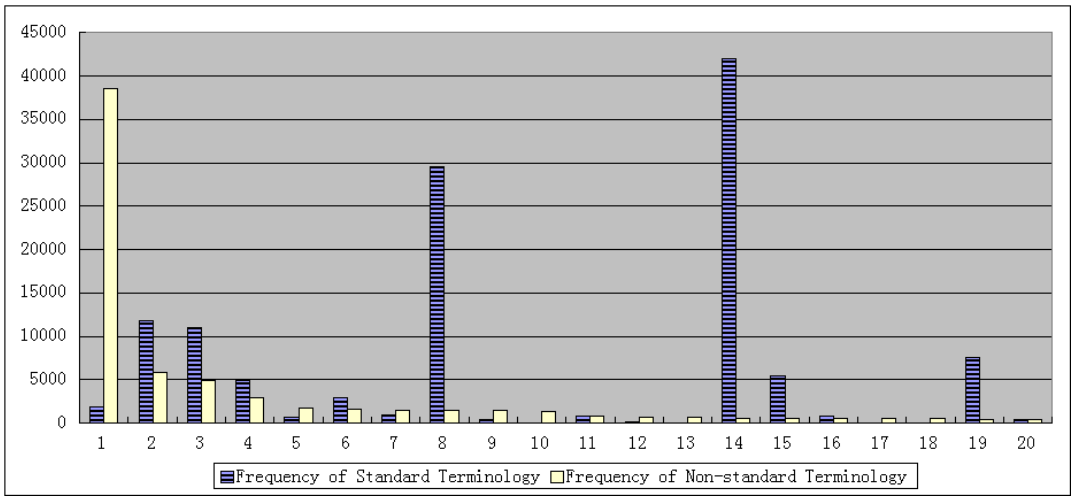

FIGURE 4. TOP 20 FREQUENCY OF NON-STANDARD TERMINOLOGY AND THAT OF THEIR CORRESPONDING STANDARD TERMINOLOGY WHICH SHARED SAME ENGLISH CONCEPT

4. **Conclusions and discussions.** To the terminology of medicine science, it is complex in the use detail of standard terminology and non-standard terminology. But at the macroscopic level, there are some conclusions, which are very like the conclusions in computer science terminology domain which we did the similar investigation in [9]:

1) The overall use of standard terminology is superior to that of non-standard terminology whereas non-standard terminology is not used by accident. From this result of situation investigation, the use frequency of standard terminology is 3 times of that of non-standard terminology.
2) To most frequent used standard terminology, the use of them are far more than that of non-standard terminology which share the same English terminology except on rare occasions.
3) In academic papers, there are 27% of non-standard terminology being used more frequent than standard terminology. It must pay special attention that, some very frequent used non-standard terminology in academic papers, are used more than that of standard terminology. It shows that the use of non-standard terminology is not by accident.

Furthermore, the use of tiny minority of non-standard terminology are predominate in the medicine science domain. Academic papers seldom use these standard terminologies when express the corresponding concept. The related institutions should study the reason and do more specific research or investigation. If using these non-standard terminology are not wrong or raise more problem, and if these situation has been going on for a long time, the better scheme of standard terminology should be updated if necessary.

Acknowledgment. This work is supported by National Language Committee Research Project (Grant No. WT125-21) and National Natural Science Foundation of China (Grant No. 61305089 and No. 61103089).

REFERENCES

[1] Richard A. Strehlow, *Standardization of Technology terminology principles and practices.* American society for testing and materials, Baltimore, 1988.

[2] Feng Zhiwei，*The origin of the Chinese terminology standardization and development.* China Standardization, 10, 2002. (In Chinese)

[3] Ma Juhong，*Some Problems and Thoughts on the Standard of Science and Technology Terms.* Terminology Standardization & Information Technology，2, 2007. (In Chinese)

[4] Liu Qing, *The basic steps of terminology standardization in brief.* Chinese Science and Technology Terms Journal，1, 2000. (In Chinese)

[5] Liu Qing, *The important role of scientific terminology normalization in information dissemination.* Chinese Science and Technology Terms Journal, 1, 2002. (In Chinese)

[6] Zhang Hui, *Standardization is a key problem of the development of science and technology.* China Terminology, 2, 2007. (In Chinese)

[7] Zeng Xiaobing, Zhang Zhiping, Liu Rong, Yang Erhong, Zhang Pu, *Investigation and Discussion on Chinese Idioms and Idiomatic Phrases in the Chinese Language Situation Report.* Journal of Chinese Information Processing, vol (22):6, 2008. (In Chinese)

[8] Liu Pengyuan, *Word Translation Disambiguation Based on Source-Target Dictionary and Target Language Corpus.* J. of Knowledge and Language Processing, Vol. 3-3; 2012.

[9] Pengyuan Liu, Yanqiu Shao and Likun Qiu. *A Situation Investigation of Chinese Computer Terminology Using in the Academic Journals.* ICIC Express Letters, Part B: Applications. Vol(6): 2, February, 2015. (Will be published)