ISCA Archive
http://www.isca-speech.org/archive

International Symposium on Chinese Spoken Language Processing (ISCSLP 2002)
Taipei, Taiwan
August 23-24, 2002

INVESTIGATION AND ANALYSIS ON DESIGNING CHINESE BALANCE CORPUS

Rile HU[1]，*Chengqing ZONG*[1]，*Juha Iso-Sipilä*[2]，*Bo XU*[1]

[1]National Laboratory of Pattern Recognition, Institute of Automation,
Chinese Academy of Sciences, Beijing，100080
{rlhu，cqzong，xubo}@nlpr.ia.ac.cn

[2]Nokia China R&D Center, Beijing，100013
juha.iso-sipila@nokia.com

ABSTRACT

Recently, the statistical methods have become the main methods in the research of computational linguistics and natural language processing. The corpus is the basis of the statistical method. How to keep the balance in corpus collection is an important issue. In this paper, we report the results of our investigation and analysis on some real corpus, and propose a scheme to keep the balance in corpus design. Suggestions for the composition in corpus design are also presented in this paper.

Key Words: Balance corpus, Corpus base design, Composition of corpus

1. INTRODUCTION

Recently, the statistical method has achieved exciting performance and is becoming more and more remarkable. It becomes the main method of natural language processing. The corpus is the basis of the statistical method. So the development of corpus base is becoming more and more important in the research of computational linguistics and natural language processing.

To keep the balance in a corpus collection is an important issue. Here, the balance means the weighting among the different sections in a general corpus. Questions associated with the balance, such as how to build a corpus base that covers a wide range of domains, are complex and often intractable. The question of balance can be totally avoided only by corpora consisting of everything published in a historical period, the complete works about a domain, or any other total population of text.[1] Obviously, balance in a corpus does not mean to have equal amounts of texts from different domains that are covered by the corpus. So, we must make effort to decide how to choose the texts reasonably from the different sources.

We have analyzed the structure of several early-built corpus bases. These corpora are: the Surrey of English Usage（SEU） Corpus, founded by Randolph Quirk in 1959, is the most important pre-electronic corpus assembled particularly for grammatical description; the Brown Corpus, collected by Brown University, is the first machine-readable corpus for linguistic research; the Longman/Lancaster English Language Corpus (LLELC), compiled at the University of Lancaster and the University of Oslo in collaboration with the Norwegian Computing Centre for the Humanities at Bergen, is a machine-readable corpus of written British English; the British National Corpus (BNC), contains both spoken texts and written texts, was designed to be well balanced[1] [3] [5] . HanYu, is a manually tagged Chinese corpus collected by Tsinghua University[2] ; Dictionary of Modern Chinese Grammar Information is a Chinese corpus compiled by Peking University [8]. These corpora have been widely used as the balance corpora.

However, about the distributions of the different domains of the corpus, no quantitative statistical works have been reported so far.

Now, the most common medias that people use to access information are Internet and newspapers. So, the research on the real texts should be based on the Internet websites and the newspapers.

In section 2, we give out our results of the investigations on changes of words in different time. In section 3, we give out our results of the investigations on distribution of corpus in different domains. In section 4, we give out our suggestions to build corpus base.

2. INVESTIGATIONS ON CHANGES OF WORDS IN DIFFERENT TIME

2.1 Motivations

We have collected two years' corpora of People Daily respectively published in 1981 and 1996. The corpora size of 1981 is 36.3M bytes, which contain 35,224,624 Chinese characters. The corpora size of 1996 is 48.6M bytes and in which there are 47,670,394 Chinese characters.

For the two years' corpora, first we segmented the words, tagged the POSs, and then counted the processed corpora and analyzed the changes of the words and the word frequencies.

2.2 Statistical Results

A. Changes of the words

There are 36,756 different words in the 1981 corpus, and 38,914 different words in the 1996 corpus.

According the statistical results, 2,257 words that are in the wordlist of 1981 corpus cannot be found in the wordlist of 1996 corpus, and these words cover 6.14% of whole wordlist of the 1981 corpus.

At the same time, 4,415 words that are in the wordlist of 1996 corpus cannot be found in the wordlist of 1981 corpus, and these words cover 11.35% of the whole wordlist of 1996 corpus.

B. Changes of the word frequencies

In order to analyze the changes of the word frequencies of these two years' corpora, we chose the top 10% most frequently used words in the two years' wordlist (T words) as the target of our investigations.

We investigated these T words, and gained the results as follows:

- ✧ Some most frequently used words are in both two wordlists. These words are the most frequently used words in our own life. Such as the words "的"、"了"、"是"、"和"、"这"、"我"and etc.
- ✧ 613 words in the top 10% most frequently used words of 1981 cannot be found in the top 10% most frequently used words of 1996. These words cover 17.15% part of the top 10% most frequently used words of 1981.
- ✧ 829 words in the top 10% most frequently used words of 1996 cannot be found in the top 10% most frequently used words of 1981. These words cover 21.30% part of the top 10% most frequently used words of 1996.

All the data of section 2.2.2 are summarized in Table-3:

Year	1981	1996
File size (MB)	36.3	48.6
Chinese characters	35,224,624	47,670,394
Words	36,756	38,914
Different words between two years	2,257	4,415
Rate of different words	6.14%	11.35%
T words	3,676	3,891
Different T words between two years	613	829
Rate of different T words	17.15%	21.30%

Table-3

2.3 Analyses on the results

1. The reason of the changes of the words in different time is due to the appearance and disappearance of the concepts and domains.

For example:

- ♦ The words "上纲上线"、"工分值" which are contained in the wordlist of the year 1981 can not be found in the wordlist of the year 1996, this is because of the disappearance of the correlative concepts.
- ♦ The words "希望工程 "、"一国两制" which are contained in the wordlist of the year 1996 can not be found in the wordlist of the year 1981, this is because of the appearance of the correlative concepts.
- ♦ The words "股市"、"多媒体" which are contained in the wordlist of the year 1996 can not be found in the wordlist of the year 1981, this is because of the appearance of the domains of the stock and IT.

2. The time's gap between the two corpora also caused changes on the word frequencies. Though some most frequently used words in our life may exist in both corpora, there are about 20% top 10% most frequently used words changes in the two corpora.

3. There are only 9 empty words in the 613 words which are in the top 10% most frequently used words of 1981 and can not be found in the top 10% most frequently used words of 1996. And there are only 10 empty words in the 829 words which are in the top 10% most frequently used words of 1996 and can not be found in the top 10% most frequently used words of 1981. These results show that the changes of the word frequencies are mostly occur on substances.

4. The changes of the most frequently used words are mostly the words in the domains of politics, commerce and science.

5. The factor of time caused the changes of the words in the two corpora are shown in figure-1:

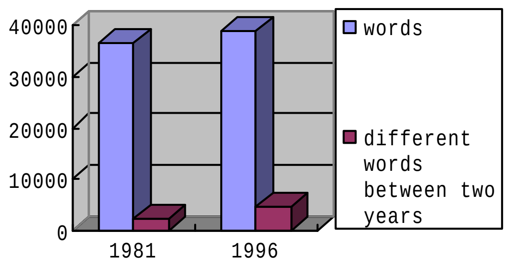

Figure-1

The factor of time caused the changes of the word frequencies in the two corpora are shown in figure-2:

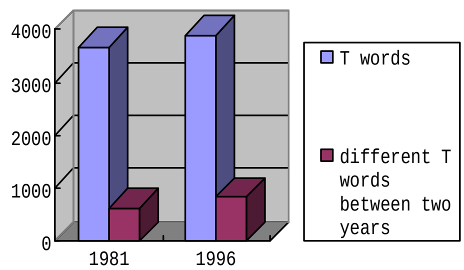

Figure-2

From the table we can see that the 15 years' gap between two corpora makes many changes on the words in the corpora. The different words between two years' corpora cover about 10% part of the wordlists. And there are also about 20% of the top

10% most frequently used words changed during the 15 years' gap. So in corpus designing, time must be considered as an important factor.

3. INVESTIGATIONS ON DISTRIBUTION OF COPPUS IN DIFFERENT DOMAIN

We have analyzed the distribution of different domains on the homepages of several famous websites written in Chinese, including Ren Min Net (www.people.com.cn), and Zao Bao Net (www.zaobao.com.sg).

A. Ren Min Net (www.people.com.cn)

We have counted all the texts on the Ren Min Net during a period of one week. The date was from March 20th, 2002 to March 26th, 2002.

There are 4,439,703 Chinese characters in these texts. All these texts cover a range of national affairs, international affairs, commerce and finance, science and education, IT, military affairs, leisure, living, sports, belief and thought, society and environment, ten domains in all. The rate of each domain is 12.1%, 10.1%, 16.8%, 4.4%, 5.8%, 5.6%, 6.8%, 9.2%, 2.8%, 9.9%, 10.4%, and 6.1%.

We summarized these ten domains into seven domains. After summarizing the distributions of the domains are shown in table-1:

Domain	Tokens	Rate
News	1,499,093	33.8%
Belief and thought	904,036	20.4%
Commerce and finance	747,987	16.8%
Science and technology	552,886	12.5%
Culture/Entertainment	300,659	6.8%
Living	409,770	9.2%
Sports	125,272	2.8%
Total	4,439,703	100%

Table-1

B. Zao Bao Net (www.zaobao.com.sg)

We have counted all the texts on the Zao Bao Net during a period of six days. The date was from March 21st, 2002 to March 26th, 2002.

There are 550,648 Chinese characters in these texts. All these texts cover a range of international affairs, IT, local affairs, belief and thought, sports, leisure, commerce and Finance, eight domains on all. The rate of each domain is 15.4%, 4.8%, 28.5%, 14.9%, 5.9%, 3.0%, 22.5%, and 5.0%.

We summarized these eight domains into six domains. After summarizing the distributions of the domains are shown in table-2:

Domain	Tokens	Rate
News	242,207	44.0%
Science and technology	26,437	4.8%
Belief and thought	81,903	14.9%
Sports	32,255	5.9%
Culture/Entertainment	16,010	3.0%
Commerce and finance	151,836	27.6%
Total	550,648	100%

Table-2

4. SUGGESTIONS TO BUILD CORPUS BASE

According to the statistical results we have gotten and the analyses on these results, we give our suggestions to build corpus base:

- ✓ We suggested that the texts you choose should be limited in 10 years. If you can choose the texts in 5 years, the representativeness of your corpus may be better.
- ✓ Our suggestions of the composition in corpus designing are shown in detail in Table-4:

Domain	Rate
News	30%～40%
Sports/Games	5%～10%
Commerce and finance	20%～30%
Culture/Entertainment	5%～10%
Living	10%～15%
Belief and thought	10%～20%
Science and technology	5%～10%
Total	100%

Table-4

The domain News may contain the subdomains of local and international affairs.

The domain Finance Commerce and finance may contain the subdomains of business, domestic and foreign market.

The domain Culture/Entertainment may contain the subdomains of music, theater, exhibitions, travel and tourism.

The domain Living may contain the subdomains of health and popular sciences.

The domain Science and technology may contain the subdomains of science, education and IT.

The domain Belief and thought may contain the subdomains of editorials, opinions and personal issues.

The rate of the domains can be chose from the bound, and you must assure that the rates of the domains sum up to a correct percentage of the overall domain.

5. ACKNOWLEDGEMENTS

The research work described in this paper is supported by the National Nature Science Foundation of China under grant number 9835003 and 60175012. It is also supported by the National Key Basic Research Program of China under grant number G1998030504, and Europe LC-Star project under grant number IST-2001-32216.

6. REFERENCES

[1] Graeme Kennedy. An Introduction to Corpus Linguistics. Foreign Language Teaching and Research Press, 2000.

[2] Feng Zhiwei. The History and Actuality of Chinese Corpus Research. In proceedings of ICCC ’ 2000, Singapore. pp. 1-15.

[3] Douglas Biber, Susan Conrad, Randi Reppen. Corpus Linguistics. Foreign Language Teaching and Research Press, 2000.

[4] Sinclair J. Corpus, Concordance, Collocation. Oxford: Oxford Press, 1991.

[5] Summers D. Longman/Lancaster English Language Corpus: Criteria and Design. Harlow: Longman, 1991.

[6] Johansson S., Stenstrom. English Computer Corpora. Berlin:Mouton de Gruyter: 33-48, 1991.

[7] Sinclair J. Reflections on computer corpora in English language research,1982.

[8] Yu Shiwen. Dictionary of Modern Chinese Grammar Information. Tsinghua University Press,1998.