

## Some Distributional Properties of Mandarin Chinese --A Study Based on the Academia Sinica Corpus

Ching-Yu Chen\*, Shu-Fen Tseng\*, Chu-Ren Huang\*\*, Keh-jiann Chen\*

\*The Institute of Information Science, Academia Sinica

\*\*The Institute of History and Philology, Academia Sinica  
Nankang, Taipei, Taiwan,  
Republic of China

### 0. Abstract

The study of word frequency has been discussed by linguists, psychologists, and computer scientists. However, the results of these studies cannot be valid unless the corpus is big enough and properly-segmented. This paper observes the distributional information derived from word frequency based on a 14-million-character corpus of Chinese newspaper (CKIP 1993). This is the first available Mandarin Chinese corpus of such magnitude. The word frequency count is obtained with an automatic-segmentation program with above 99% accuracy rate (Chen and Liu 1992). The count reflects some general phenomena of Chinese usage. For example, among the first thousand high frequency words, there are more bi-syllabic words than mono-syllabic words, attesting to the trend of bi-syllabicification observed by many linguists. However, in general, the mono-syllabic function words occur more frequently than bi-syllabic words. In addition, the frequency of numerals is ranked according to their numeric order ('one' is higher than 'two', and 'two' is in turn higher than 'three', etc.)

This paper discusses the theoretical and applicational implications of these distributional properties. For instance, we find that the most frequent 2452 characters and 28124 words make up 99% of the corpus content. It is suggested that the optimal strategy for learning Chinese lies in the mastery of the most frequent 2452 characters plus words whose meanings can not be predicted on the basis of their component characters. This implies that one need not know 28124 words in order to achieve good reading knowledge in Chinese. Given the noted parallel between the internal structure of words and phrases, one can predict that knowledge of a few thousand words and of the morphosyntactic rules will enable one to read Chinese without much difficulty.

### 1. Overview

Previous studies on word frequencies were not based on large corpora. For example, Hsieh (1975) studied word frequency based on Taiwan's seven leading daily newspapers, which contained a corpus of only 112,708 words. In addition, Hsieh's work was done by hand and not automatized, so there might have some miscalculation in the result. Beijing Language College's (1985) 'Xiandai Hanyu Pinlu Cidian' (Word Frequency count of Modern Mandarin), a well-known dictionary which is often cited, has 1,808,114 words. However, the result of these studies cannot be valid unless the corpus is big enough and properly-segmented. This paper observes the distributional information derived from word frequency based on a fourteen-million-character corpus of Chinese newspapers (Huang et al, 1993), (Huang and Chen, 1992). This is the first available Mandarin Chinese corpus of such magnitude. The word frequency count is obtained with an automatic-segmentation program with above 99% accuracy rates. (Chen and Liu, 1992). Furthermore, since the corpus contains mostly texts from journals, its contexts cover many topics, such as politics, humanities, sciences, culture, arts and literature....etc. It also contains interviews, fiction, letters....etc. In other words, this corpus has both critical size and

diversity. The distributional properties that obtain from the corpus should be a good indicator of the general properties of Mandarin Chinese.

In this study, we follow approaches in statistical linguistics and try to combine mathematics and linguistics in our research. Through observing computed results, we are able to gain an overall understanding of the distributional properties of languages. In section 2, we will make observations based on the word frequency count, and discuss the linguistic interpretation of these observations. In section 3, we provide statistics derived from our frequency count to test the robustness of some important laws proposed in the field. In the last section, section 4, we will make some concluding remarks on this study.

## 2. The Linguistic Phenomena and Study

In this section, linguistic phenomena are observed and interpreted.

### 2.1 Classification of the 500 Most Frequent Words

The first 500 words occur no less than 2778 times. These words (types) make up 50.696 percentage of the corpus. There are some important attributes of these most frequent words:

(1) Among the 500 most frequent words, there are 93 disyllabic nouns, and many of them are government organizations, corporations, and official titles (32 nouns): *zheng fu* 'government', *xian fu* 'county government', *guo jia* 'nation', *li wei* 'legislator', *yi yuan* 'councilman', *xian zhang* 'county magistrate', etc.) These words are all frequently used words in political news.

(2) Among the first 500 most frequent words, there are 136 verbs, and the active verbs are more than stative verbs (84:50), transitive verbs more than intransitive verbs (99:35). Among disyllabic verbs the frequency of discourse verbs is comparatively high. For example *biao shi* 'to express', *zhi chu* 'to point out', *ren wei* 'to think', *jue ding* 'to decide', *bao dao* 'to report', *diao cha* 'to investigate', *gui ding* 'to prescribe'....etc, and for the most part action verbs occur with single objects. Among 99 transitive verbs, there are 57 action verbs with single objects.

(3) In addition, since the three factors, "person, place and time", are the three (almost) obligatory elements in literal actions or states, they are also the most common properties of the first five hundred high frequency words. For example, there is the factor of "person", and as we mentioned before, most of them are government organizations and official titles. The factor of "time" includes words such as: *mu qian* 'presently', *zuo tian* 'yesterday', *jin nian* 'this year', *shang wu* 'morning', *qu nian* 'last year', etc. The factor of "place" including *Tai Wan* 'Taiwan', *Tai Bei* 'Taipei', *Mei Guo* 'America', *Ri Ben* 'Japan', *Kao hsiung* 'Kaohsiung', etc. also occurs frequently.

As mentioned above, among the first five hundred high frequency words, there are 93 disyllabic nouns, and 32 of them are names of government organizations, corporations and official titles while there are also 12 time nouns and 24 place nouns. These three kinds of words make up of two thirds disyllabic nouns.

## 2.2 Distribution of Syllabic Length

Table 2-1 is computed based on a corpus of 9,529,233 segmented words. Segmentation was done by the automatic-segmentation program designed by Chinese Knowledge Information Processing Group (Chen & Liu, 1992). The numbers of words and frequency of one-character words to nine-character are given in Table 2-1.

Concerning word type, there are 5191 monosyllabic words, which consists of 9.52% of all lexical entries. There are 35,752 disyllabic words and they consist of 65.60% of all entries. The numbers of trisyllabic and quarter-syllabic words are very close (12.36% and 11.58% respectively). Words of five or more characters are rare, about 0.94%. However, concerning word tokens, the numbers of monosyllabic words is more than the numbers of disyllabic words (53.77% vs. 42.28%). The sum of the two classes of tokens is more than 96%, while the other words which are more than three characters only add up to less than 4%.

From the statistics, we can see that most Mandarin Chinese words are monosyllabic or disyllabic. The pre-dominance of disyllabic word types (65.60%) seem to support the theory that Chinese is in the process of disyllabification. However, in actual use monosyllabic words are far more frequently than disyllabic words. Moreover, we count the average word length of Mandarin Chinese is 1.494 according to the table; which is lower than the estimated value of 2.

Kind of Word	Number of Words	Total Frequency	Type	Token
One-character words	5191	5123836	9.52%	53.77%
Two-character words	35752	4028894	65.60%	42.28%
Three-character words	6736	279711	12.36%	2.94%
Four-character words	6309	91006	11.58%	0.96%
Five-character words	300	3635	0.55%	0.04%
Six-character words	138	1736	0.25%	0.02%
Seven-character words	58	337	0.11%	0.00%
Eight-character words	15	72	0.03%	0.00%
Nine-character words	1	6	0.00%	0.00%
<b>Total words</b>	<b>54500</b>	<b>9529233</b>	<b>100.00%</b>	<b>100.00%</b>

Table 2-1 Words Classified by Syllabic Length

We will next investigate more closely the distribution in terms of word-length by monitoring the

distribution of each 100 word segments on the frequency scale. The result is Figure 2-1.

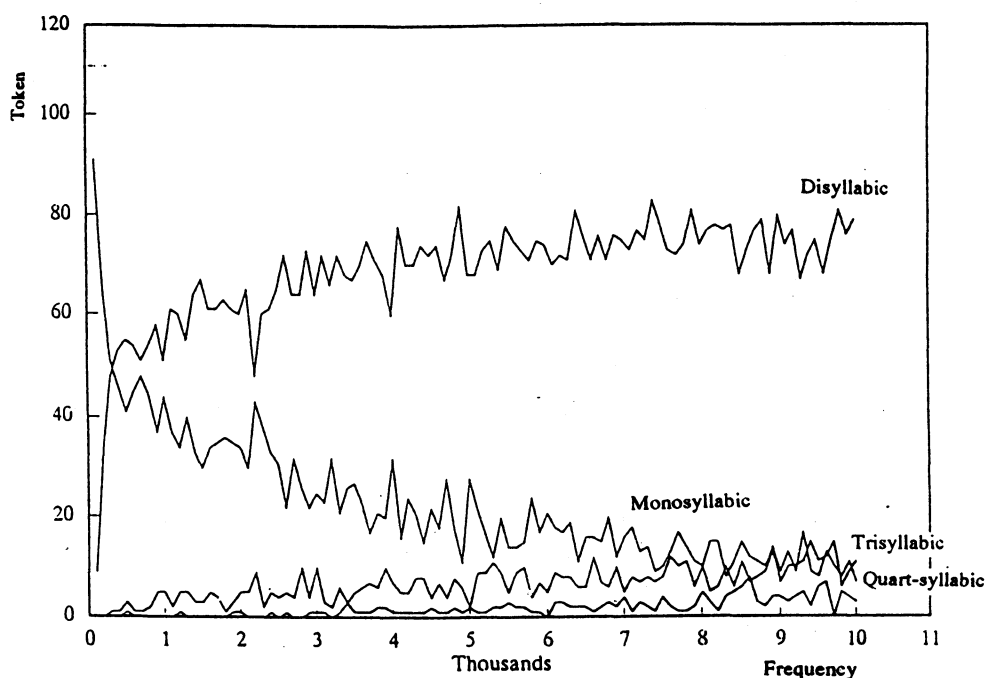


Figure 2-1, Distribution of Word Types with regard to syllable number within each 100-word frequency stage

Among the 200 most frequent words, there are no multisyllabic words longer than three syllables. Moreover, in the 300 most frequent words, monosyllabic words are far more than disyllabic words (monosyllabic: 156, disyllabic:44). The numbers of disyllabic words overtakes the number of monosyllabic words in the 300-400 stage. From Figure 3-1 we can see that with the 300 most frequent words, monosyllabic and disyllabic words show dramatic decrease and increase respectively. Then from the 300th words to the 10000th words, the count of monosyllabic continues to decrease, whereas disyllabic words are increasing continuously. Because longer multisyllabic words consist only a small percentage, the two curves of monosyllabic and disyllabic words in figure 3-1 are almost perfect mirror image of each other. This again shows that most Chinese words are either monosyllabic or disyllabic.

In addition we learn that the total frequency of one to four character words reaches 99%, and five and more-character words are rare. After observing the spread of every one to four-character words, we find the one-character words are predominant in the highest frequency range, and most of the words are function words such as prepositions, determinative, measures, conjunctions, personal pronouns, the verb "to be," and the verb "to have." In the next highest frequency range (400 to 2000), two-character words are predominant, and most of the words are nouns and verbs. Almost all three and four-character words are nouns and verbs. Focusing on the phenomenon, we would discuss in 3.3 why one-character function words have such a high usage frequency. In addition, the distribution of one to four-character words in terms of grammatical categories will also be discussed.

### 3.3 High Frequency of One-Character Minor Category Words

Among high frequency words, monosyllabic words dominate, and these monosyllabic words are almost all minor category words, which include prepositions, determinative, conjunctions, personal pronouns, etc.. Of all monosyllabic words, *de* has the highest frequency. Next we will observe the distribution of

prepositions, determinative, and conjunctions.

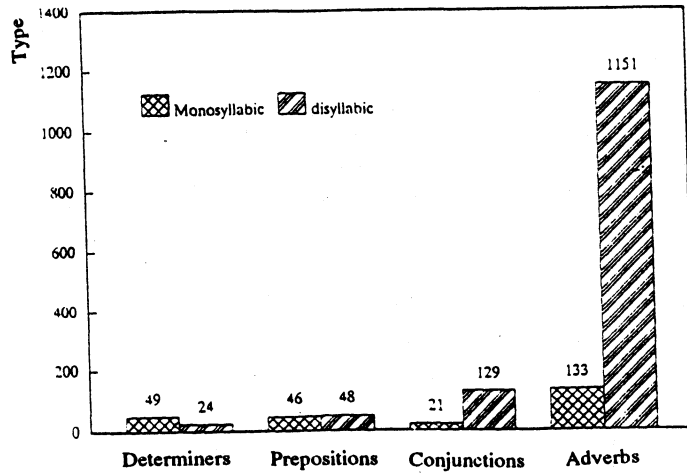


Fig.2-2 the distribution of minor category words in Mandarin Chinese (bar graph)

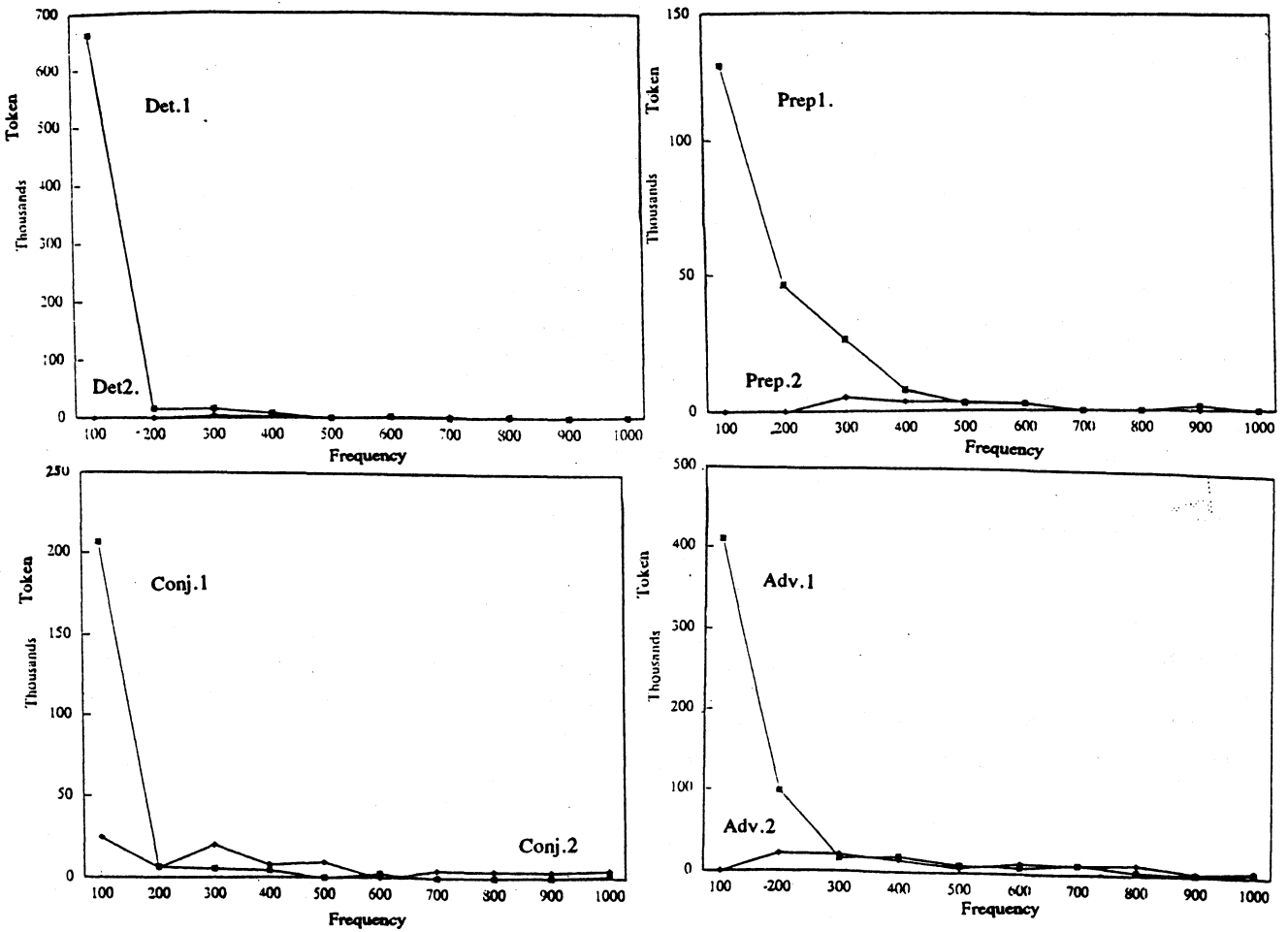


Fig.2-3 the distribution of minor category words in Mandarin Chinese (line graph)

In Figure 2-2, we see the number of one-character determinatives is double of that of two-character determinatives. In Figure 2-3, the graph shows 20 one-character determinatives appear in first 100 high frequency words, which occupy half of the total amount of one-character determinatives, but two-character determinatives do not appear before the 200th word. Thus we see one-character determinatives are greater in number and also in usage frequency. The amount of one-character prepositions are almost equal to that of two-character prepositions, and in first 1000 words, two-character prepositions appear less than one-character prepositions. The presentations of conjunctions and adverbs also clearly illustrate the phenomenon that one-character words have higher frequency than two-character words, but in first 1000 words, the amount of two-character words' appearances are few. Hence, two-character conjunctions and two-character adverbs are all low-frequency words.

Since Chinese is generally not inflectional, it is necessary to use functional category words to represent grammatical relations, thus they occupy an important position in the grammar as well as use. But these words have low productivity and belong to a closed class. So the chance of repetitive use is very high. The obligatoriness of functional category words, such as having no proforms and allowing no ellipsis, explain the reason why one-character function words occupy a majority in instances of high frequency words. In addition to the discussion of Fig.2-2 and Fig.2-3 above concerning the distribution of function words, we make detailed observations on these words and find the following phenomena:

(1) In the first 1000 words, many one-character words are ranked higher than two-character words which have the same meaning: conjunctions *ji* and *yi ji* 'and', *chie* and *er chie* 'and', *yin* and *yin wei* 'because', *dan* and *dan shi* 'but', prepositions *zi* and *zi cong* 'since', *ju* and *gen ju* 'according to', and *dui* and *dui yu* 'toward', for instance. Maybe it presents the characteristics of the writing form that writing vocabulary is necessary to be brief and clear, simple and to the point to save the space of printing plate. Besides, since function words only have syntactic function, if one-character words do work, we must refrain from using two-character words, so that we can avoid verbiage.

(2) It is important to take 'syllable' into considerations when using Chinese, especially when choosing adequate adverbs to modify some verbs. It is observed that some monosyllabic adverbs always occur with some certain monosyllabic verbs. Since these verbs are frequent, the frequency of these adverbs are also very high. According to our corpus, high frequency verbs (*shi* 'to be', *you* 'to have', for example) always occur with adverbs (*jiang* 'be going to', *bu* 'not', *ye* 'also', *yi* 'already', *dou* 'all', *ying* 'should', *zai* 'not yet', for example.) These adverbs also have high frequency.

#### 2.4 Distribution of Major Categories

In one to four-character words, the distribution of noun frequency and verb frequency have some differences in addition to similarities. The similarities are in the distribution of noun frequency and verb frequency: the frequency of monosyllabic words is higher than that of disyllabic words, and the frequency of disyllabic words is higher than those of three and four-character words. The differences lies in the frequency rank of four-character words. Three and four-character nouns occur in the set of the 500 most frequent words. Hence, in three and four-character words, the usage frequency of nouns is higher than that of verbs. However, multi-syllabic words do not rank higher than 2500th and four-character verbs do not rank higher than 4500th.

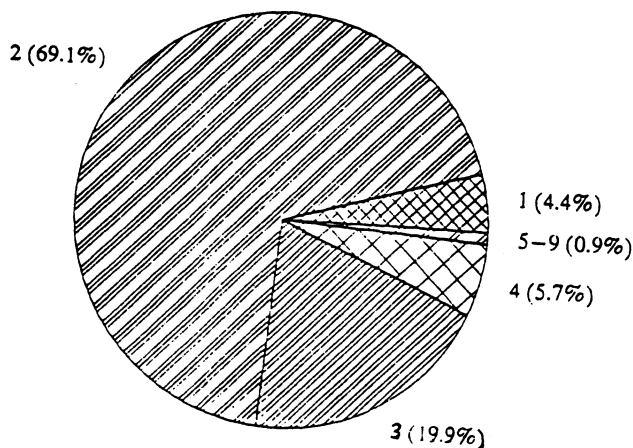


Fig. 2-4 ratio chart of Noun types in Mandarin Chinese

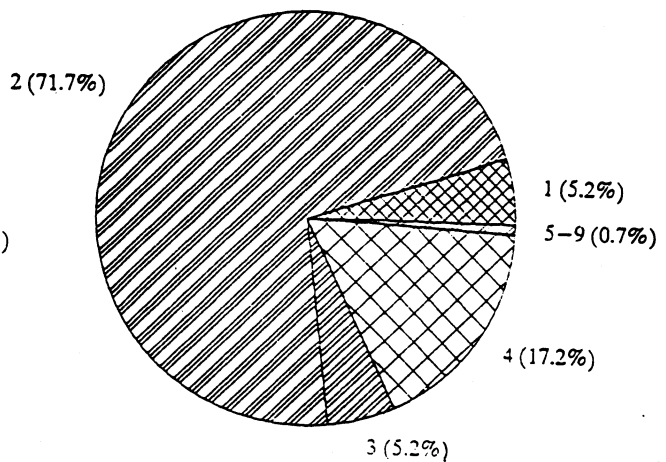


Fig. 2-5 ratio chart of Verb types in Mandarin Chinese

In addition, we can see from the ratio chart (Figure 2-4 and Figure 2-5) the percentage of every syllable types of nouns and verbs---the ratio of one and two-character nouns and verbs are similar, but that of three and four-character words are contrary; unexpectedly there are more three-character nouns but more four-character verbs.

Based on the data of corpus-based frequency count of words(CKIP, 1993), three-character nouns are mostly derived words, i.e., words composed of stems and affixes. These words have the often refer to government institution(--*Yuan*, --*Yu*, --*Shu*, etc.), name of administration division(--*Shi*, --*Xian*, etc.). Because of the high-productivity, three-character nouns consist a significant percentage. Chinese names in general consist of three-characters; this may be one of the reasons why there are many three-character nouns.

Four-character nouns are almost always proper names and government corporations, but four-character government corporations are usually abbreviated to disyllabic words (for example, *Zhong Yang Yin Hang* --> *Yang Hang* 'Central Bank'). As a result, four-character nouns occur less often, and their frequency is not high. To sum up, except for monosyllabic words, the amounts of nouns reduce progressively as the characters increase.

The distribution of three and four-character verbs is different to that of nouns. There are a few three-character verbs, which are almost VR compound verbs (*ying xiang dao* 'influence', *bian geng wei* 'change') and V-O construction verbs (*da dian hua* 'to telephone', *fa pi qi* 'to lose temper'). In four-character verbs, VR compound verbs are few, and most of them are idioms (cheng2 yu3). As is well-known, four character *Cheng2 Yu3* is the time-honored way to conventionalize and lexicalize longer expressions in Chinese. Since these idioms are often used to creat vivid speech, four-character verbs are more than three-character verbs.

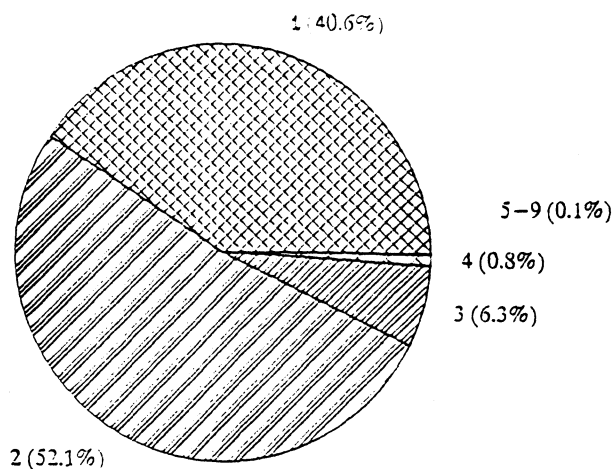


Fig.2-6 ratio chart of Noun tokens in Mandarin Chinese

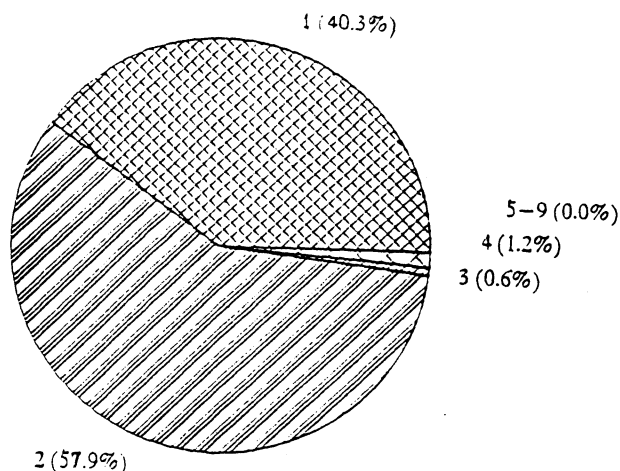


Fig.2-7 ratio chart of Verb tokens in Mandarin Chinese

Figure 2-6 and Figure 2-7 are the ratio chart of tokens of nouns and verbs based on syllabic length. Comparing Figure 2-4 and Figure 2-6, we see that the percentage of monosyllabic nouns expands to nearly ten times (4.4% type vs. 40.6% token) when we use them, but three and more-character words correspondingly contract (e.g. for 3 character words, percentages come down to 6.3% from 19.9%). To see the rise and fall of verbs, we see that the extension degree of monosyllabic verbs is equivalent to that of nouns, but the usage frequency of three and more-character words reduces more drastically (only 1.8%). We learn from Figure 2-6 and Figure 2-7 the main present forms of nouns and verbs are one and two-character. The reason why three-character nouns still occupy a significant ratio is that nouns are designator of entities and cannot be easily abbreviated without causing ambiguities. In contrast, three and more-character verbs occupy only 1.8%, because they do not show strong negative effect when abbreviated. As to the reason why four-character verbs are more than four-character nouns, it is because there are many idioms (Cheng<sub>2</sub> Yu<sub>3</sub>) in Chinese which can be used as predicates, but, in fact, in contrast with Figure 2-5, the type amount of four-character verbs occupy 17.2%, and it contracts to 1.2% when being actually used. We see that the frequency of four-character idioms is not high in common usage, though they represent a healthy portion of the lexicon.

### 3.5. Another Distributional Property: numerals

All the fundamental numerals one to ten occur among the most 50 highest frequent words. Their frequencies generally reflect the numeric order, except for *wu* 'five' and *shi* 'ten'.

In the corpus, the high frequency of numerals is related to their common use in counting and referring. The progressive decrease from one to nine can be explained by some characteristics we meet when using ordinal numbers to count. In our statistics of words which display numerals side by side, we find a large quantity of numerals are used along with standard measures ("dollar", "year", "month" and "day", for example) and quasi-measures for measuring place words (*xiang* 'alley', *nong* 'lane', and *hao* 'number', for example.)

When we use ordinal numbers to refer a group of things, the range to number would influence the usage frequency of every number. For example, in a year we just have "the first season" to "the fourth



season", so the numerals of five and above are not used in this context and consequently occur less frequently. Thus, the frequencies of numerals from one to nine usually decrease gradually.

The reason for one's highest frequency is predictable, because "one" covers many meanings. In Chinese, besides the meaning "number", it also presents the meaning "whole" and "same". The abbreviation of the frequencies of "five" and "ten", exceptional to numeric order, relate to the system we use to count. The numbers over ten would usually have the number "ten" in them, "five" has a higher frequency than "four" probably because "five" is the middle value of "ten" and we are used to generalize the number less than five with "five" (for example, we always say "about 25 dollars" instead of "23 dollars"). The importance of the number 5 in Chinese can also be supported by the idiom *Yi Wu Yi Shi* '(literally) per-five, per-ten', '(idiomatically) to give a detailed account', and the fact that Chinese abacus uses both decimal and quintuple units.

## 2.6 Abbreviation

The efficiency concern of modern life also reflects on human language. People use abbreviation more and more frequently; we can easily observe the phenomena in the corpus. For example, with the same meaning, *guo min da hui dai biao* (nation-people-grand-meeting-representative) 'the National Assembly' is less frequent than its abbreviation *guo da dai biao*, whereas, *guo da dai biao* is in turn less frequent than its abbreviation *guo dai*. Predictably, abbreviation words are found among the most frequently used words. For example, *Yang Hang* 'Central Bank of China', *Tai Da* 'National Taiwan University'. We find among the 2500 most frequent words.

In addition, the syllabic transformation of abbreviations and their origin forms are interesting. We found that words with odd syllables in its full forms are most likely to be abbreviated to odd syllable ones. Whereas the even syllable words are abbreviated to even syllable words. It is rare that some trisyllabic words are shortened to disyllabic words. We only find counterexamples to this generalization in the title of a news story, such as *Jing Bu* (shortened from *Jing Ji Bu* 'Ministry of Economic Affairs'), *Li Yuan* (shortened from *Li Fa Yuan* 'Legislative Yuan'). This can again be used as evidence to support the generalization that people use abbreviated form for the sake of efficiency but do not sacrifice their communicative goals.

## 3. An Observation on Statistics Linguistics — Zipf's Law

It is claimed that when we arrange the result of word frequency count in a decreasing order, it happens that the rank multiplies the rate of its frequency results in a constant; i.e.  $F * R = C$  (R: rank, F: the rate of frequency) This is known as Zipf's Law. (Zipf, 1949) Following Zipf's proposal, there was a lot of discussion on it in the literature. However, our work is different from previous studies in some aspects:

1. Our study is based on a much larger corpus than previous ones; their research was based on at most a few hundred-character corpus.
2. This is the first time Zipf's Law is applied in Chinese with a properly segmented words. Previous work focused their research on Chinese character frequency instead of word frequency.

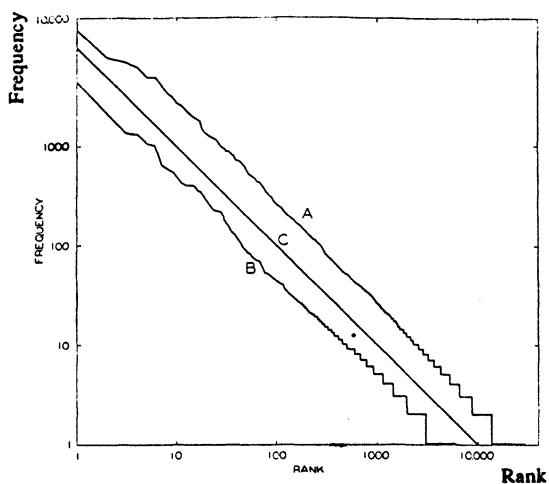


Fig.3-1 the rank-frequency distribution of words (Zipf, 1949)

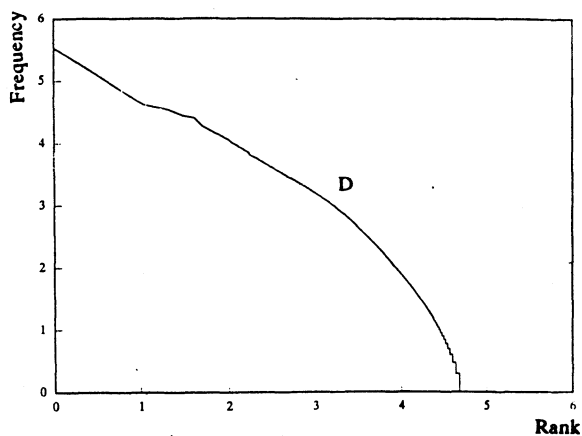


Fig.3-2 the rank-frequency distribution of words (Academia Sinica Corpus)

Firstly, the rank-frequency distribution of words (Zipf, 1949) is shown in Fig.3-1; Curve A is the James Joyce data; B the Eldridge data; C ideal curve of  $45^\circ$  slope. Curve A and Curve B are close to a straight line. In Fig.3-2, curve D, shows the rank-frequency distribution of words derived from Academia Sinica corpus. We can see the curve approximates linear between 42th and 1408th. This follows Zipf's prediction. Scholars (Deng, 1987) have claimed that Zipf's Law can not apply the most frequent words and the rare frequent words, so the the curve in Fig.3-2 does not violate the spirit of Zipf's Law.

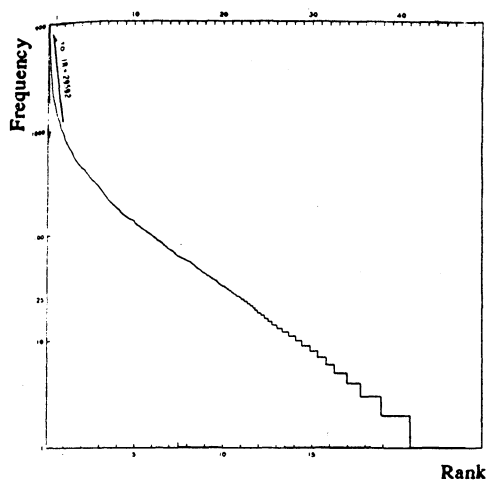


Fig.3-3 the rank-frequency distribution of Chinese characters (Zipf, 194)

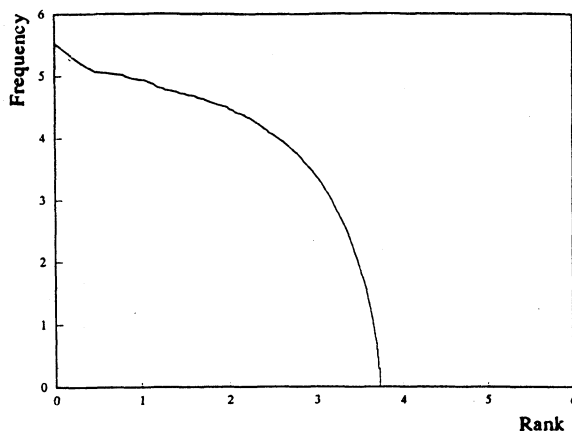


Fig.3-4 the rank-frequency distribution of Chinese characters (Academia Sinica Corpus)

Besides, Fig.3-3 and Fig.3-4 demonstrate the rank-frequency distribution of Chinese characters. Fig.3-3 is Zipf's data; and Fig.3-4 based on Academia Sinica corpus. The curve of Fig.3-3 is firstly downwardly convex then becomes linear and finally becomes step-like. However, Fig.3-4 shows a upwardly convex curve while Fig.3-3 is downwardly convex. The difference between these two figures implies that Zipf's Law does not correctly predict the distribution of Chinese characters. The reason

might be that not all Chinese characters are information units.<sup>1</sup>

Thus, the distribution of a corpus is more complex than what Zipf predicted, and it is possible that Zipf's Law can only fit a part of a corpus, not whole corpus. And if Zipf's Law can apply the distribution of characters should be reconsidered. Thus whether the value of C is 0.1 should not be emphasized as previous studies do.<sup>2,3</sup>

In conclusion, we have shown that Zipf's Law can not be a general property of the distribution of Chinese characters. However, it still applies to some specific range of word distribution. The interpretation given in Smith (1991) should shed light on why Zipf's Law applies in a limited domain: "It may suggest an equilibrium between unwillingness to exert mental energy in coming up with words and the need for words specific enough to express the meaning. Or it may suggest that, as an efficient channel of communication, language obeys laws of probability by the number of available word choices."

#### 4. Conclusion

All the above discussion and observation are based on the CKIP word frequency count which is computed from the Academia Sinica Corpus. Our research provides empirical evidence which lend solid ground to linguistic theory and prediction. In addition to providing empirical evidences to linguistic theory, our research also captures distributional properties of Chinese that cannot be predicted by pure theoretical approaches. For instance, although 5665 Chinese characters in total occur in the 14-million-character corpus, the frequently used 2452 characters made up 99 percentage of the corpus. This figure implies that a person who has learned 2452 Chinese characters plus a few morphological rules can easily understand most of a Chinese texts. The result can suggest an expected scale for the evaluation of Chinese learners (native and foreign).

In conclusion, this study suggests a new approach combining computer and linguistic theory. In Taiwan, this is the first time the frequency count of words is directly analyzed and observed on a completely electronically based corpus. With the success of this pioneering corpus-based study of Chinese linguistics, more extensive utilization of corpuses in linguistic and NLP research should bear profitable results in the future.

#### Acknowledgements:

Research for this project was partially funded by the Chiang Ching-Kuo Foundation for Internation Scholarly Exchanges. We want to thank Guan-Wen Wang, Sheng-Yih Wang, Wei-Liang Chen, Spring Ji, Wen-Shyang Lu for their programming and drawing graphs. We also want to express our gratitude to Kathleen Ahrens for her comments on the earlier draft and to all the colleagues at CKIP for providing support and information. Any remaining errors are our responsibility.

#### References

[1] The Chinese Knowledge Information Processing Group, 1993, "Corpus-Based Frequency

- Count of Words in Journal Chinese", The CKIP Group, Institute of Information Science, Academia Sinica, Taiwan, ROC.
- [2] The Chinese Knowledge Information Processing Group, 1993, "Corpus-Based Frequency Count of Characters in Journal Chinese", Academia Sinica, Taiwan, ROC.
- [3] The Chinese Knowledge Information Processing Group, 1993, "The Most Frequent Verbs in Journal Chinese and Their Classification", Academia Sinica, Taiwan, ROC.
- [4] The Chinese Knowledge Information Processing Group, 1993, "The Most Frequent Nouns in Journal Chinese and Their Classification", Academia Sinica, Taiwan, ROC.
- [5] Chu-Ren Huang and Keh-jiann Chen, 1992, "A Chinese Corpus for Linguistic Research". In the Proceedings of the 1992 International Conference on Computational Linguistics (COLING-92). 1214-1217. Nantes, France.
- [6] Chu-Ren Huang, et al., 1993, "Chinese Linguistic computing---Modern and Classical Chinese Corpora at Academia Sinica". In the Proceedings of Fifth ROC-Japan Information Symposium on Modern Information Services. 67-93. Taipei, Taiwan. R.O.C.
- [7] George W. Smith, 1991, "Computers and Human Language", Oxford University Press
- [8] Hsieh, Jen Hsiao, 1975, "A Frequency Count of Contemporary Chinese Vocabulary Based on Seven Leading Newspapers", South Carolina University, Ph. D., U.S.A.
- [9] John R. Pierce, 1980, "An Introduction to Information Theory --Symbols, Signals & Noise", Dover Publications, Second, Revised Edition, New York, U.S.A.
- [10] Keh-jiann Chen, Shing-Huan Liu, 1992, "Word Identification For Mandarin Chinese Sentences". In Proc. of COLING 92, Nantes, France.
- [11] Zipf, 1949, "Human Behavior and the Principle of Least Effort", Addison-Wesley Press, Massachusetts, U.S.A.
- [12] 尹斌庸 (Yin, Bin-Yung): 漢字習得效率研究, <<語文建設通訊>>第38期, 1992.12
- [13] 王還..等 (Wang, Huan et al.): 現代漢語頻率詞典(上,中,下) 北京語言學院出版社, 1985.7
- [14] 衛志強 (Wei, Zhi-Qiang): 當代跨科學語言學, 北京語言學院出版社, 1992.2, 第一版

[15] 鄧洛華 (Deng, Lou-Hua): 詞頻分析, <<武漢大學學報>>(社會科學版)第一期, 1987

[16] 馮志偉 (Fong, Zhi-Wei): 數理語言學, 知識出版社, 上海, 1985.8

Footnote

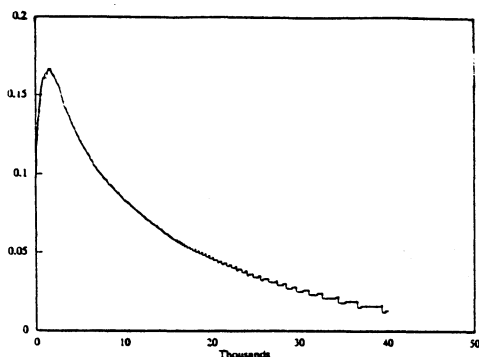


Fig.3-5 value of C in Academia Sinica Corpus

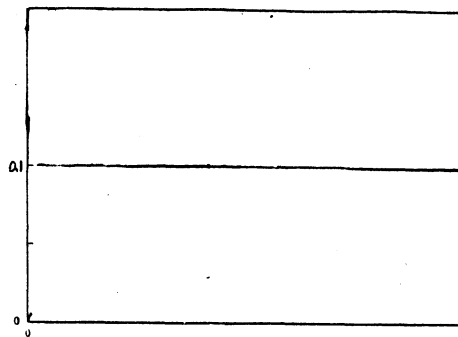


Fig.3-6 the Zipf's ideal value of C

1. It can also be observed that Pierce's(1980) account "...Cree gives a line having only about three-fourth the slope of the Zipf's law line. This means a greater number of different words in a given length of text --- a large vocabulary. Chinese characters give a curve which zooms up at the left, indicating a smaller vocabulary" is misleading. Chinese does not have a smaller vocabulary. It has a relatively small set of characters.

2. Let us examine Zipf's formula the way previous scholars did. The following numbers are the products of F multiplying R. All the product approximate to 0.1 like previous researchers' results.

0.035 <sub>1</sub>	0.024 <sub>2</sub>	0.032 <sub>3</sub>	0.030 <sub>4</sub>	0.031 <sub>5</sub>
0.046 <sub>10</sub>	0.075 <sub>20</sub>	0.100 <sub>50</sub>	0.120 <sub>100</sub>	0.125 <sub>200</sub>
0.162 <sub>1000</sub>	0.163 <sub>2000</sub>	0.147 <sub>3000</sub>	0.133 <sub>4000</sub>	0.121 <sub>5000</sub>
0.111 <sub>6000</sub>	0.102 <sub>7000</sub>	0.096 <sub>8000</sub>	0.090 <sub>9000</sub>	0.083 <sub>10000</sub>

However, when we continue computing the products of F multiplying R and plotting the result on a XY diagram (see Fig.3-5), the graph is a mountain-like curve which zooms up at the upper left and suddenly drops off to the lower right. Also, we can see that the curve peaks at the rank of 1378. This can be compared with the idealized prediction based on Zipf's Law. (Fig.3-6)

It is interesting that our curve is plotted between the reasonable range that Zipf predicted, but it shows a smooth curve and the value of C seems to be relative to its rank.

3. In fact, Fong(1985) has proved that Zipf's Law is just a specific condition of Mandelbrot's formula. There are three variables in Mandelbrot's formula. This implies it existing at least three conditions should be controlled in such experiment. Thus the universality of Zipf's Law should be reconsidered.

Appendix

Words Token Frequency

的一十是有三爲將五及人不中之以四也而與日這他六六七時八已後個會百對並但年上元表九於等月該由多萬和各說公所兩陳都第被可因其此則能分就林至到目來	334639 113149 101384 71824 59375 55913 54768 48256 44740 43421 41938 40043 38352 37697 37678 37254 36881 36807 36699 35974 20 34528 33532 33301 32741 30368 30004 29918 28159 28148 28105 27755 27716 27108 27025 26993 26759 26596 25922 25796 25694 40 25550 25482 25116 24255 23940 23499 22638 21885 21872 19050 19012 18869 18193 18163 17995 17130 17065 16891 16741 16503 60 16468 16104 15786 15528 15216 15114 14921 14806 14725 14665 14582 14322 14194	3.513 4.701 5.766 6.520 7.143 7.730 8.305 8.812 9.281 9.737 10.178 10.598 11.001 11.396 11.792 12.183 12.570 12.957 13.342 13.720 20 14.082 14.434 14.784 15.128 15.446 15.761 16.076 16.371 16.667 16.962 17.253 17.544 17.829 18.112 18.396 18.677 18.956 19.228 19.499 19.769 20 20.037 20.305 20.568 20.823 21.074 21.321 21.559 21.788 22.018 22.218 22.418 22.616 22.807 22.997 23.186 23.366 23.545 23.723 23.898 24.072 60 24.245 24.414 24.579 24.742 24.902 25.061 25.217 25.373 25.527 25.681 25.835 25.985 26.134	下內千要次應指出 名向高警昨天未地再由前又台或仍因小廿新每最項張種台問家卻者使更案經外我全市黃認天國位路曾歲李鄉人單組政大無點才工自市均股他們億今年即她業者	14078 14050 14032 13983 13948 13849 13760 80 13477 13336 13217 12956 12772 12749 12599 12510 12440 12376 12277 12275 12024 11989 11957 11940 11711 11518 11506 11494 100 11431 11370 11118 10875 10812 10586 10473 10241 10135 10111 10085 10059 9999 9981 9978 9886 9854 9662 9607 9597 120 9588 9502 9453 9307 9282 9275 8995 8928 8841 8823 8813 8739 8649 8605 8548 8515 8505 8458 8341 8287 140 8268 8259 8227 8202 8107	26.282 26.429 26.577 26.723 26.870 27.015 27.160 27.301 27.441 27.580 27.716 27.850 27.984 28.116 28.247 28.378 28.508 28.637 28.766 28.892 29.018 29.143 29.269 29.392 29.513 29.633 29.754 29.874 29.994 30.110 30.224 30.338 30.449 30.559 30.667 30.773 30.879 30.985 31.091 31.196 31.300 31.405 31.509 31.612 31.714 31.815 31.915 32.016 32.116 32.215 32.313 32.410 32.508 32.602 32.696 32.789 32.881 32.974 33.066 33.156 33.247 33.336 33.426 33.515 33.604 33.691 33.778 33.865 33.952 34.038 34.124 34.210	王北行 區亦可正民衆有日 很從處較要 吳國望 美希得如著起還事縣號車卅內 文依讓有關區者 投資 且發工生程 活間決發發成餘人許據鎮金無另表過 報導 本場部做此些省好用方調影響把發過社會	8063 8034 7988 7965 7830 7819 7770 7768 7746 7737 7611 7590 7566 7502 7409 160 7401 7399 7390 7274 7266 7261 7224 7196 7136 7102 7101 7088 7043 6989 6781 6738 6711 6641 6472 6438 180 6363 6348 6321 6317 6309 6213 6210 6204 6179 6172 6130 6125 6122 6103 6102 6074 6040 5992 5986 5975 200 5968 5966 5959 5887 5840 5828 5823 5808 5786 5772 5741 5723 5722 5672 5652 5635 5634 5622	34.294 34.379 34.462 34.546 34.628 34.710 34.792 34.873 34.955 35.036 35.116 35.196 35.275 35.354 35.432 35.509 35.587 35.665 35.741 35.817 35.893 35.969 36.045 36.120 36.194 36.269 36.343 36.417 36.491 36.562 36.633 36.703 36.773 36.841 36.908 36.975 37.042 37.108 37.174 37.241 37.306 37.371 37.436 37.501 37.566 37.630 37.695 37.759 37.823 37.887 37.951 38.014 38.077 38.140 38.203 38.265 38.328 38.390 38.452 38.514 38.575 38.636 38.697 38.758 38.818 38.878 38.939 38.999 39.058 39.118 39.177 39.236 39.295	中心明 提出以生方濟 提可學地經隊只日本時果成 如造成若女華達某警上午於 性法因爲線是國去美自看長南國際在定 劉水中共約方 局是午 但下出姓相其中他 其令土地比參加 建雖然 給數廠商我們幾 業包時間請使發進來 黨未遭分	5617 5604 5588 5580 5573 5570 5534 5527 5470 5445 5440 5344 5337 5322 5316 5304 5304 5288 5231 5219 5193 5175 240 5141 5138 5125 5123 5122 5114 5104 5092 5045 5043 4980 4974 4971 4947 4918 4902 4884 4859 4853 4850 260 4829 4806 4802 4797 4797 4753 4723 4723 4706 4705 4690 4673 4662 4627 4624 4612 4596 4590 4569 4532 280 4531 4518 4512 4483 4473 4469 4452 4446 4434 4416	39.354 39.413 39.471 39.530 39.588 39.647 39.705 39.763 39.820 39.878 39.935 39.991 40.047 40.103 40.159 40.214 40.270 40.325 40.380 40.435 40.490 40.544 40.598 40.652 40.706 40.760 40.813 40.867 40.921 40.974 41.027 41.080 41.132 41.184 41.237 41.289 41.340 41.392 41.443 41.494 41.545 41.596 41.647 41.697 41.747 41.798 41.848 41.898 41.948 41.997 42.047 42.096 42.145 42.194 42.243 42.292 42.340 42.389 42.437 42.485 42.533 42.581 42.628 42.676 42.723 42.770 42.817 42.864 42.911 42.958 43.004 43.050
--	--	--	---	--	--	--	---	--	---	---	--

立委	4416	43.097	委員會	3642	46.115	通過	3158	48.711	屈了解	2755	50.928
立去清	4351	43.142	研究	3634	46.153	個人	3137	48.744	了住	2752	50.957
必須	4350	43.188	企業	3626	46.192	學校	3134	48.777	準備	2750	50.986
代舉	4345	43.234	近	3611	46.229	學價	3126	48.810	據據	2749	51.015
行成	4342	43.279	男	3609	46.267	最後	3123	48.842	涉已	2744	51.044
成爲	4339	43.325	份	3595	46.305	---440---			查獲	2744	51.072
回	4337	43.370	行	3593	46.343	打	3122	48.875	行	2743	51.101
計	4330	43.416	動	3585	46.380	事	3110	48.908	極	2741	51.130
畫	4309	43.461	決	3584	46.418	件	3104	48.940	院	2738	51.159
---300---	4276	43.506	分	3579	46.456	周	3102	48.973	政	2737	51.187
那	4251	43.551	十	3569	46.493	言	3091	49.006	積	2733	51.216
德	4245	43.595	道	3568	46.531	永	3089	49.038	中	2730	51.245
檢	4241	43.640	嚴	3551	46.568	男	3088	49.070	平	2729	51.273
加	4233	43.684	官	3548	46.605	謝	3086	49.103	---520---		
國	4223	43.729	至	3547	46.642	環	3085	49.135	利	2726	51.302
辦	4221	43.773	共	3543	46.680	保	3077	49.167	用	2721	51.331
理	4217	43.817	中	3542	46.717	術	3072	49.200	員	2720	51.359
方	4214	43.861	國	3532	46.754	具	3058	49.232	校	2719	51.388
場	4165	43.905	安	---380---		電	3054	49.264	校	2712	51.416
難	4154	43.949	山	3519	46.791	話	3045	49.296	香	2710	51.445
情	4153	43.992	建	3514	46.828	是	3036	49.328	港	2707	51.473
連	4126	44.036	設	3514	46.865	意	3035	49.360	份	2702	51.501
行	4097	44.079	件	3503	46.901	度	3034	49.391	開	2699	51.530
原	4095	44.122	市	3499	46.938	舉	3031	49.423	展	2696	51.558
段	4078	44.164	安	3498	46.975	戶	3029	49.455	電	2693	51.586
除	4077	44.207	全	3497	47.011	報	3024	49.487	視	2692	51.615
上	4069	44.250	別	3491	47.048	非	---460---		週	2691	51.643
盤	4066	44.293	信	3470	47.085	他	3020	49.519	北	2687	51.671
進	4054	44.335	條	3468	47.121	規	2989	49.550	獲	2675	51.699
部	4051	44.378	天	3468	47.157	劃	2979	49.581	得	2667	51.727
銀	---320---		量	3460	47.194	總	2963	49.612	乃	2666	51.755
行	4050	44.420	政	3458	47.230	統	2954	49.643	心	2656	51.783
長	4027	44.463	策	3452	47.266	境	2953	49.674	蘇	2655	51.811
要	4019	44.505	能	3439	47.302	中	2945	49.705	移	2649	51.839
主	4013	44.547	計	3428	47.338	龍	2944	49.736	---540---		
半	4012	44.589	3427	47.374	派	2938	49.767	類	2643	51.867	
海	4002	44.631	形	3418	47.410	費	2928	49.798	舉	2638	51.894
期	3995	44.673	對	3410	47.446	需	2905	49.828	場	2628	51.922
東	3989	44.715	營	3406	47.482	作	2905	49.859	式	2621	51.949
非	3978	44.757	行	---400---		業	2902	49.889	興	2615	51.977
配	3954	44.798	產	3389	47.517	伊	2899	49.920	運	2606	52.004
合	3953	44.840	品	3374	47.553	拉	2897	49.950	業	2606	52.031
員	3953	44.881	近	3364	47.588	克	2897	49.980	金	2605	52.059
議	3940	44.922	通	3358	47.623	北	2897	50.011	市	2604	52.086
服	3931	44.964	增	3334	47.658	台	2894	50.041	加	2601	52.113
務	3931	44.964	請	3330	47.693	市	2889	50.072	至	2599	52.141
會	3930	45.005	申	3323	47.728	北	2887	50.102	甚	2599	52.168
關	3911	45.046	費	3314	47.763	加	---480---		設	2596	52.195
係	3898	45.087	日	3286	47.797	甚	2878	50.132	去	2594	52.223
理	3897	45.128	高	3285	47.832	想	2861	50.162	圖	2587	52.250
立	3894	45.169	而	3285	47.866	提	2860	50.192	整	2582	52.277
治	3886	45.210	局	3276	47.901	往	2857	50.222	所	2579	52.304
政	---340---		引	3272	47.935	政	2857	50.252	圾	2578	52.331
治	3874	45.250	重	3265	47.969	績	2856	50.282	計	2577	52.358
們	3859	45.291	僅	3264	48.004	繼	2851	50.312	放	2577	52.385
之	3843	45.331	強	3263	48.038	公	2850	50.342	---560---		
後	3832	45.371	如	3262	48.072	走	2835	50.372	持	2568	52.412
先	3814	45.411	何	3258	48.106	支	2833	50.401	千	2568	52.439
結	3813	45.451	農	3247	48.141	持	2830	50.431	萬	2567	52.466
果	3797	45.491	近	3244	48.175	社	2826	50.461	反	2560	52.493
嫌	3788	45.531	雙	---420---		波	2813	50.490	會	2552	52.520
村	3783	45.571	方	3236	48.209	興	2804	50.520	主	2551	52.546
太	3775	45.610	定	3226	48.242	重	2800	50.549	管	2549	52.573
開	3768	45.650	了	3225	48.276	另	2800	50.579	券	2548	52.600
始	3750	45.689	除	3212	48.310	外	2798	50.608	轉	2538	52.627
榮	3743	45.729	里	3209	48.344	生	2791	50.637	取	2536	52.653
獲	3742	45.768	權	3208	48.377	共	2783	50.667	上	2535	52.680
出	3714	45.807	廠	3205	48.411	只	2778	50.696	團	2535	52.706
所	3713	45.846	商	3203	48.445	美	---500---		開	2533	52.733
以	3694	45.885	站	3200	48.478	國	2778	50.725	料	2528	52.760
及	3673	45.923	化	3177	48.512	機	2777	50.754	光	2527	52.786
訊	3671	45.962	講	3173	48.545	實	2772	50.783	任	2521	52.813
當	3664	46.000	受	3170	48.578	施	2765	50.812	何	2520	52.839
生	---360---		低	3167	48.611	教	2765	50.841	級	2520	52.865
見	3660	46.039	國	3165	48.645	入	2761	50.870	義	2518	52.892
係	3658	46.077	小	3158	48.678	頭			取	2518	52.918
同			少			道			街		
開			加			路			任		
發			利			預			主		
提			化			算			直		
供			中								
許											
多											
蔡											