Reading news for information:
How much vocabulary a CFL learner should know

Jun Da
Middle Tennessee State University
jda@mtsu.edu

Abstract: This paper reports the findings of a corpus-based study on the vocabulary used in journalistic Chinese. Based on a 20-million character corpus of more than 27,000 news texts collected between mid 2003 and the end of 2004 from various Chinese media sources in different countries and regions, a character frequency list and three word and phrase frequency lists with two, three and four-characters were compiled. With reference to a consolidated vocabulary list that is based on six manually edited word and phrase lists in the public domain and contains more than 140 thousand entries, we estimate that the number of two- to four-character words and phrases used in journalistic Chinese is between 80,000 and 90,000. With the information about vocabulary size and usage in journalistic Chinese, we suggest that CFL learners should first focus on around 20,000 high to medium frequency words and phrases if they want to develop basic ability in Chinese news reading. Further, around 70,000 low frequency vocabulary is needed if they want to develop full proficiency in Chinese news reading.

摘要：本文报道一个关于新闻汉语所用词汇的研究结果。本研究使用了一个二千万字的汉语新闻语料库，其中收集了从2003年中至2004年底由世界各地多家中文媒体发表的各种题材的时事新闻报道2万7千多篇。我们根据该语料库建立了一个单字频率表和三个分别由二字，三字和四字组成的词和短语频率表。根据对六个在互联网上公开的总数为14万多条的词汇表的检索对比结果，我们推测新闻汉语所用的词汇总数约在8万和9万之间。基于我们对新闻汉语所用词汇的计算和推测结果，我们建议母语为非汉语的汉语学习者要获得基本的新闻阅读能力应该首先重点学习2万个左右的高中频词汇。如果要获得完全的汉语新闻阅读能力，他们则需要进一步学习或知晓7万左右的低频词汇。

1. Introduction

Reading involves identifying words in texts and forming an understanding from them. Past research indicates that vocabulary knowledge is the single most important factor contributing to reading comprehension (c.f., for example, Laflamme 1997). In descriptions of readers' vocabulary knowledge as well as their effect on reading comprehension, researchers often refer to the number of words that native speakers know and the number of words needed to do the things that a foreign language learner needs to do (Nation and Waring 1997).

In the case of English, for example, it has been found that educated native English speakers know about 16,000 to 20,000 word families (Goulden, Nation and Read, 1990; Zechmeister, Chronis, Cull, D'Anna and Healy, 1995), where a word family is defined as a headword, its inflected forms and its closely related derived forms (from affixation, etc.) (Nation 2001). Further, research has indicated that while readers can decipher the meaning of words as they read, adequate reading comprehension depends on a person already knowing a very high percentage of words found in any give texts. In his experiment on the easiness of reading texts in relation to vocabulary knowledge of English native speakers, for example, Carver (1994) found that in easy reading materials nearly 0% of the words will be unknown to his subjects. In difficult materials around 2% or more of the words will be unknown. With appropriate reading materials, around 1% of the words will be unknown to his subjects. Those observations suggest that 99% coverage rate is needed for pleasure reading of difficult materials.

In the case of non-native speakers reading in English, Laufer (1992, as cited in Nation and Waring 1997) suggests that a vocabulary of 3,000 word families of general English is enough for a good understanding of general English texts such as a novel. Based on information provided in Hirsh and Nation (1992), Nation and Waring (1997) put the number between 3000 and 5,000 word families. Waring and Nation (2004) claim that the number of words needed for the reading of technical texts such as science texts and newspapers is larger than for less formal texts. In terms of vocabulary coverage rate, Liu and Nation (1985) and Laufer (1989) suggested a level of 95% for adequate reading comprehension. Hu and Nation (2000) found that unless there is at least 98% or higher coverage rate of the running words in a text, the probability of successful guessing of unknown words will be severely reduced.

As compared with the availability of numerous studies on the size and coverage rate of English vocabulary necessary for adequate reading comprehension for both native and non-native speakers, our knowledge about similar issues in Chinese is less conclusive. One cause of the problem is that our estimates on vocabulary size in Chinese could vary greatly depending on whether we rely on information from manually compiled dictionaries or the amount of vocabulary reported in empirical studies on Chinese textbooks or prescribed for CFL (Chinese as a Foreign Language) learning. In the case of manually compiled dictionaries, for instance, The Unabridged Chinese Dictionary (《汉语大词典》) contains more than 370,000 entries that include characters, words, phrases and idiomatic expressions. In contrast, medium-sized dictionaries such as Modern Chinese Dictionary (《现代汉语词典》) that are intended for daily use by educated native speakers contain between 50,000 and 60,000 entries of characters, words, phrases and idioms, etc.

However, some empirical studies have put the amount of vocabulary in real language use on a smaller scale. For example, a study on Chinese textbooks used in both elementary and middle schools in mainland China conducted by the Modern Education Technology Research Institute, Beijing Normal University[1] in the 1990s found that out of the 704,841 words identified, only 39,601 are unique. In a similar study on Chinese textbooks used in both elementary and middle

[1] c.f. Hui Wang. Statistical studies on Chinese vocabulary (王慧：汉语词汇统计研究). Online version at <http://www.huayuqiao.org/articles/wanghui/wanghui06.doc>. Last checked: 2005-06-20. The date of publication is unknown from the online source.

schools in mainland China, researchers at Beijing Language and Culture University[2] found that the first 1000 words with the highest frequencies make up more than 74% of the total words identified where the average word length is 1.98 characters[3].

In the case of CFL, vocabulary lists compiled for learning and instruction has put the number on an even smaller scale. For instance, the vocabulary list compiled by the China National Office for Teaching Chinese as a Foreign Languages (国家对外汉语教学领导小组办公室《汉语水平词汇与汉字等级大纲》, henceforth the HSK vocabulary list or HSK list) contains 8,882 characters, words and phrases.

Such a great variation in the estimates of vocabulary size in Chinese is often accompanied by the scarcity of empirical information on vocabulary use in real language situations, especially in the public domain. Thus, it is difficult, if not impossible, for researchers to rely on those estimates to conduct research on vocabulary and reading comprehension of CFL (Chinese as a Foreign Language) learners.

In this paper, we report the findings of a corpus-based research project that is intended to provide a partial solution to the above mentioned problem. The project has two main objectives: First, by using both automatic method and a manually compiled word and phrase list, we are going to build a comprehensive list of words and phrases used in current news reports in Chinese that are made up of two, three and four characters. Secondly, based on statistical information from our word and phrase lists as well as six other manually edited vocabulary lists available in the public domain, we will provide an estimate on the vocabulary size necessary for CFL learners who are interested in developing proficiency in reading current news from various Chinese news media across regions. We choose news texts as the focus of our study based on the belief that news reading is a typical case of reading for information that requires a larger number of words than reading in less formal texts (such as novels). Empirical knowledge about the vocabulary and its size used in current Chinese news reports will give us a clear picture about the upper limit of vocabulary size in other subject areas.

In the past, similar studies based on corpus of Chinese news texts have already been conducted. For example, researchers at the Hong Kong Polytechnic University conducted a study between 1991-1997 on a 6-million character corpus containing news articles collected between 1990 – 1992 from newspapers published in mainland China, Taiwan and Hong Kong. Their Chinese Word Bank from Mainland China, Taiwan and Hong Kong (《中国大陆、台湾、香港汉语词库》) contains 60,811 entries[4]. Further research by Chen and Tang (1999)[5] on the word bank identified 12,700 frequently used words and found that the three regions share a common collection of high- and medium-frequency words that makes up 90% of the total number of words identified and covers 95% of the text materials. The remaining 10% words that are not shared among the three regions concentrate on the low-frequency range. It is also reported in Feng (2002) that several universities in mainland China such as Peking University and Beijing Language and Culture University conducted researches on corpora based on materials from People's Daily and

[2] Formerly known as Beijing Institute of Language.

[3] c.f. Footnote 1.

[4] c.f. Footnote 1.

[5] 陈瑞端、汤志祥. 1999. 九十年代汉语词汇地域分布的定量研究. 语言文字应用. 3,10-18. Cited in Hui Wang (c.f. Footnote 1).

Xinhua News Agency. Unfortunately, the findings of those studies are not made available in the public domain.

This project differs from previous similar studies in two major aspects. First, our news corpus contains more recent news texts from more regions and media sources, which is in contrast to previously known empirical studies (such as the one conducted at Hong Kong Polytechnic University) that were conducted in the 1980s and (early) 1990s and based on limited selection of news materials. Given the constant changing nature of news contents and news reporting, we believe that results from our study are more appropriate for CFL reading research and instruction at the current time. Secondly, in this study we will rely on both automatic method and previously manually edited vocabulary lists to identify words and phrases found in our corpus and will not take context information into consideration when we generate lists of words and phrases. This is in contrast to several previous studies (as citied in our introduction) that relied on manual segmentation of words and phrases in running texts. While some information (such as a word's different senses of meaning) may be lost in this statistical approach to word segmentation, we believe that other information such as vocabulary size and frequency distribution obtained in this study will still be useful for researches and learning instruction in Chinese.

2. This project

2.1 Word and phrase in written Chinese

In inflectional languages such as English, individual words in running texts are delimited by white spaces. Statistical studies on the vocabulary in those languages make distinctions among tokens (a count of every word in a text), types (unique words in a text), lemma (a headword and some of its inflected and reduced forms) and word families (a headword, its inflected forms and its closely related derived forms from affixation, etc.) (Nation 2001).

However, running Chinese texts do not use delimiters (white space) or other linguistic means (such as markers) to separate individual characters, words or phrases from each other. At the same time, a word or phrase in written Chinese could contain one, two, three or even more characters. Hence, without context information, it is extremely difficult (if possible at all both in theory and practice) to use an algorithm to category character strings such as 下雨 (*to rain*, or *it's raining*) and 吃饭 (*to have meal*, or *eat rice*), etc. as a word or phrase.

In this research, no distinction is made between words and phrases. Instead, our interest is on those two-, three- or four-character sequences that are possible candidates for words or phrases in Chinese. In this sense, our use of the term vocabulary in the rest of the paper roughly corresponds to the concept of types used in the study of inflectional languages (c.f., Nation 2001).

2.2 The news corpus

The Chinese news texts used in this study were collected between the middle of 2003 and the end of 2004 from the current news collection of the World Forum website (世界论坛网) at http://www.wforum.com/gbindex.html. While a small portion of the texts were retrieved manually, the majority of the texts were harvested automatically with a computer script. The corpus contains a total of 27,965 news articles that fall into 15 different categories, as shown in Table 1.

Table 1 List of categories and number of articles in the news corpus

Categories	Number of articles
Commentary	997
Culture and education	560
Economy and finance	1580
Entertainment	693
Headline news	4726
Hong Kong and Macao	695
International	3767
Mainland China	2652
Military and defense	2547
North America	2586
Overseas Chinese	439
Science and technology	877
Social	1936
Sports	1249
Taiwan	2661
Total	**27,965**

A random sampling of those news articles indicates that the majority of the news texts originated from different news agencies or media sources in Mainland China, Taiwan, Hong Kong, Macao, other Asian countries or regions, North America, and Europe that include, but not limited to 新华社, 凤凰卫视, 中央社, 路透社, 美联社 and 世界日报, etc. A small portion of the news commentaries were written by individuals and published on the Internet. To the best judgment of the author, those news texts are mostly news summaries that were created originally in Simplified Chinese, or converted from Traditional Chinese or translated from other languages (such as Japanese) by the website news editor(s).

It should be pointed out that while the selection of those news articles may have been influenced by the preferences of the website news editor(s), we believe that the great diversity in their sources, regions and subject areas nevertheless makes them a good representative of journalistic Chinese that are currently used world wide.

2.3 Precompiled word and phrase list

In order to facilitate the identification of words and phrases from the news corpus, a consolidated word and phrase list was compiled from six manually edited vocabulary lists that are available in the public domain, which include the HSK vocabulary list[6], the Chinese-English

[6] Electronic version of the HSK vocabulary list was retrieved from http://www.chinese-forums.com/vocabulary/ on 2005-05-25. While the official count is 8.882, the online version contains 8,743 entries.

Dictionary (CEDICT)[7] database, consolidated word and phrase list by Adrian Robert[8], the 1985 word frequency list compiled by Beijing Language and Culture University[9], the word and phrase list from the Chinese Lexical Analysis System created by Institute of Computing Technology, Chinese Academy of Science (ICTCLAS)[10] and the word and phrase list from Richwin[11]. Details of the six word and phrase lists are shown in Table 2.

Table 2 Consolidated word and phrase list based on six online sources

Character/words /phrases	HSK	CEDICT	Robert	Word85	ICTCLAS	Richwin	Consolidated
Single character	1866	6851					
Two-character	6373	12944	23167	11014	43164	73396	82532
Three-character	306	2686	3692	636	17877	19411	31965
Four-character	188	1983	2651	682	9287	25868	30806
More than four characters	10	563	487	14	465	1654	2529
Subtotal	**8743**	**25027**	**29997**	**12346**	**70793**	**120329**	**147832**

In Table 2, HSK refers to the HSK vocabulary list; Word85 refers to the word frequency list compiled by Beijing Language and Culture University in 1985; and Consolidated refers to the word and phrase list consolidated by the author based on the six word and phrase lists. For the sake of convenience in discussions in this paper, we will refer to the consolidated word and phrase list as Da's consolidated list.

2.4 Data processing

Computing tasks carried out in this study were performed on the FreeBSD platform using both Unix Shell commands and customized scripts written in the PHP scripting language . MySQL is used as the backend database software.

2.4.1 Pre-processing

All files retrieved from the World Forum website are HTML-encoded webpages, which contain not only news texts but also menus and webpage footers in Chinese that are automatically added as part of the webpage templates of the website. Before character frequencies and n-grams were

[7] CEDICT was created by Paul Denisowski and is currently maintained by Erik Peterson. Data from CEDICT was retrieved from http://www.mandarintools.com/cedict.html on 2005-05-20.

[8] Robert's consolidated list was retrieved from http://kamares.ucsd.edu/~arobert/chinese_f.html on 2005-05-25.

[9] Word frequency list from Chinese Word Frequency Statistics and Analysis (《汉语词频的统计和分析》) by Beijing Language and Culture University (formerly Beijing Institute of Languages) was retrieved from the Chinese Pinyin and Input Method Forum (〖汉语拼音与输入法论坛〗) at http://sh.netsh.com/bbs/1951/. The list was posted by Fengzi (冯子) on 2002-01-11.

[10] Information about ICTCLAS (中科院计算所汉语词法分析系统) can be found at http://mtgroup.ict.ac.cn/~zhp/ICTCLAS/index.html. Vocabulary data incorporated in our consolidated list was retrieved from http://download.pchome.net/php/dl.php?sid=12405 on 2005-05-20. The original vocabulary data are intended for automatic segmentation of Chinese words and phrases in running texts and may contain some entries that are portions of a word or phrase.

[11] Word and phrase list from Richwin was retrieved from http://technology.chtsai.org/wordlist/duoyuanpinyin.zip on 2005-04-30. The list is intended for Chinese input used in the Richwin system and hence may contain entries that are portions of words or phrases.

counted, those HTML tags as well as webpage menus and footers (e.g., 编辑：龙芯 and 世界论坛网, etc.) were removed with customized PHP scripts.

2.4.2 Character segmentation

In this study, we used computer scripts written in PHP to automatically segment individual characters in running Chinese text. Since continuous character strings in running written Chinese are not delimited with any white spaces, the reliability of any automatic character segmentation depends on the encoding scheme found in the text files.

To our best knowledge, webpages at the World Forum website are encoded in the GB13000 (or GBK) standard with a two-byte encoding scheme. However, a random sampling of news articles in our corpus indicates that there are a very small portion of files that contain either Traditional Chinese characters or Japanese texts. They are most likely due to operation errors made by the website news editors when they converted news texts from their original sources (such as in Traditional Chinese) into Simplified Chinese. In our computing, those non-Simplified Chinese characters were not removed, because we believe that they represent real language use and their quantity is very small and hence will not significantly affect the outcome of this study. Instead, each character candidate was checked against the GBK character table in the character segmenting process.

2.4.3 N-gram counting

Previous research has shown that monosyllabic and disyllabic words are the majority in Chinese[12]. In this study, we counted n-grams that contain two, three or four characters. An n-gram is defined as a text string with more than one consecutive characters. To count n-gram frequencies, we used a modified method based on Brew and Moens (2000): All running Chinese texts were first divided into segments of continuous character strings where both GB encoded symbols (such as punctuation marks) and ASCII codes (e.g., white space characters) were treated as delimiters. N-grams were then identified and counted within each continuous character string.

It should be pointed out that n-grams are a superset of multi-character words or phrases in Chinese. For example, it is possible that a bigram is a two-character word, part of a word containing more than two characters, or simply a senseless random combination of two characters. Previous studies on word collocation have shown that Mutual Information value is a good measure of the strength of association between two elements in a bigram, especially when raw frequency counts of individual bigrams are high (>5) (c.f., Church and Mercer 1993). Accordingly, we computed Mutual Information values for all bigrams found in this study so that they can be used as a reference for the identification of disyllabic words or phrases. In addition, Mutual Information values are calculated for each two-character pair in a trigram. The two Mutual Information values can, for example, be used to judge if the trigram can be taken as a word or phrase in Chinese[13].

[12] c.f. Footnote 1.

[13] It may be more appropriate to calculate a generalized mutual information value for any given trigram (c.f., for example, Magerman, D.M. and M. P. Marcus. 1990. Parsing a natural language using mutual information statistics. Proceedings of the 8th National Conference on Artificial Intelligence. 984—989. Boston, MA).

2.5 Results

In this section, we present summary statistics about character and n-gram frequency distributions based on the news corpus. Detailed lists of characters and n-grams and their frequency distribution information are made available at http://lingua.mtsu.edu/chinese-computing/newscorpus/.

2.5.1 Character frequency distribution

From those 27,965 news articles in our corpus, a total of 22,256,047 characters were counted and 6,364 unique characters identified. Their cumulative distribution is shown in Tables 3 and 4.

Table 3 Cumulative number of characters in terms of percentages

10%	25%	50%	75%	90%	95%	98%	99%	99.5%	100%
7	38	155	419	838	1204	1742	2184	2651	6364

Table 4 Cumulative frequency distribution in terms of individual characters

100	500	1000	1500	2000	2500	3000	3500	5000
40.6%	79.2%	92.3%	97.0%	98.7%	99.4%	99.7%	99.8%	100%

Two observations about the characters and their distribution can be made. First, characters used for foreign proper names such as 斯, 基, 克 and 尔, etc. appear as high frequency characters. This pattern is expected, since the news corpus contains a rather large portion of international news reports. Secondly, a random inspection of very low frequency characters in the list and the original texts from which they were identified shows that some of them (in the neighborhood of 100) are due to errors made by our automatic character segmentation scripts. Their presence in the list is unwanted but unavoidable, since without manual assistance, any automatic methods for Chinese character segmentation will not be 100% fool-proof. Caution is thus suggested when information concerning those very low frequency characters is used elsewhere.

2.5.2 Bigram distribution

A total of 891,047 bigrams were counted, whose raw frequencies range from 1 to 118,299. The list of bigrams is checked against Da's consolidated list of words and phrases and 62,065 are found to be present in the list. Table 5 lists bigram frequency distribution in different frequency ranges. Note that in order to facilitate discussions later in this paper (c.f., Section 3.2), we also counted the number of words and phrases in our bigram list that are also present in the HSK vocabulary list.

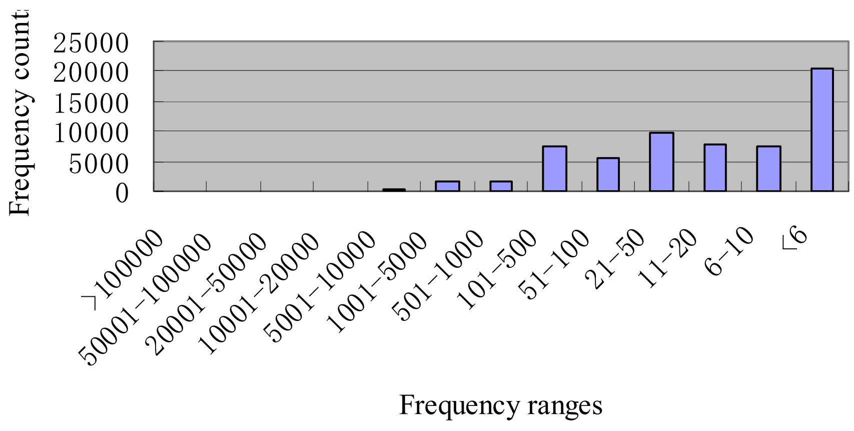
Table 5 Bigram frequency distribution

Frequency range	Raw frequency	In Da's consolidated list	In the HSK list
>100000	1	1	0
50001-100000	2	2	1
20001-50000	18	16	14
10001-20000	76	71	56
5001-10000	242	205	147
1001-5000	2185	1505	956
501-1000	2691	1511	702
101-500	19960	7568	2145
51-100	21767	5666	854
21-50	52125	9705	792
11-20	62995	7895	352
6-10	88249	7557	163
<6	640776	20363	145
Total	**891087**	**62065**	**6327**

In terms of all the bigrams generated in our computing, we find in Table 5 that the majority of bigrams (about 71.9%) are in the very low frequency range (X≤5), most of which are presumably not meaningful words or phrases. Further, medium to low frequency (6≤X≤500) bigrams range from around 20,000 to 90,000 and make up about 27.5% of the total bigrams identified. In contrast, high frequency bigrams (X>500) make up only a very small portion (around 0.6%) of the total bigrams identified.

However, different distribution patterns are observed in those bigrams that are identified as also present in Da's consolidated list, where high-frequency bigrams (X>500) make up 2.9%; medium to low frequency (6≤X≤500) 61.9% and very low-frequency (X≤5) bigrams 32.8%. Figure 1 provides a visual representation of the frequency distribution of those bigrams found to be present in Da's consolidated list.

Figure 1 Bigrams also present in Da's consolidated list



Among those high-frequency bigrams present in Da’s consolidated list, we find not only functional words such entries as 可能, 可以, 这个, 但是, and 什么 etc. but also a high concentration of proper names such as 中国[14]，美国，日本，北京，大陆 and 布什, etc.

2.5.3 Trigram distribution

A total of 4,290,983 trigrams were counted, whose raw frequencies range from 1 to 19,741. The list of trigrams is checked against Da’s consolidated list of words and phrases and 17,301 are identified as present in the list. Table 6 lists trigram frequency distributions among different frequency ranges, including the number of trigrams found to be also present in the HSK vocabulary list.

Table 6 Trigram frequency distribution

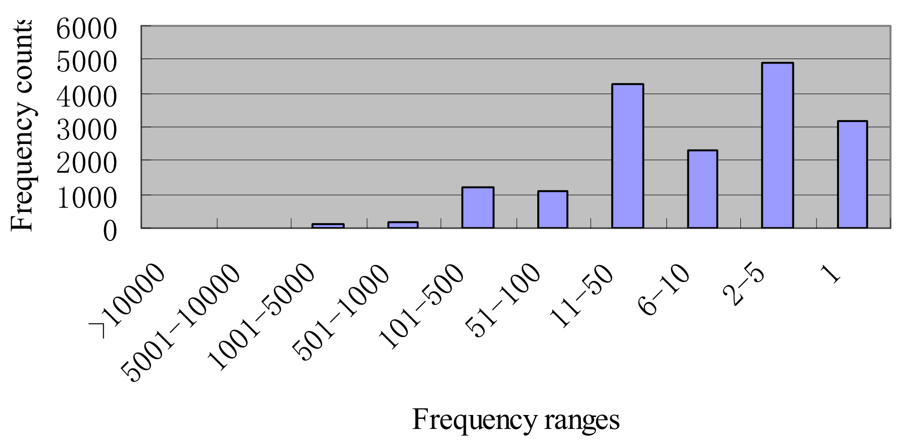
Frequency range	Raw frequency	In Da’s consolidated list	In the HSK list
>10000	4	3	0
5001-10000	13	7	0
1001-5000	469	129	22
501-1000	1147	189	15
101-500	14136	1203	67
51-100	21820	1102	39
11-50	186538	4268	72
6-10	215799	2319	23
2-5	1212187	4917	33
1	2638870	3164	9
Total	**4290983**	**17301**	**280**

In Table 6, we find that high-frequency trigrams again make up a small portions of all the trigrams and very low frequency (X≤5) trigrams the majority. Presumably the majority of those very low-frequency trigrams are random sequences of characters.

However, among those trigrams identified to be present in Da’s consolidated list, a different pattern is observed. While high-frequency trigram still makes up a small portion, the number of trigrams are more or less distributed evenly between medium to low frequency (6≤X≤500) and very low frequency (X≤5) ranges, as shown in Figure 2. A random inspection of those trigrams indicates that most of them are nouns or noun phrases and proper names such as 伊拉克, 陈水扁, 胡锦涛, 联合国 and 新华社 stay at the top of the trigram list.

[14] It remains a mystery why 中国 is not included in the HSK vocabulary list.

Figure 2 Trigrams also present in Da's consolidated list



2.5.4 Quadrigram

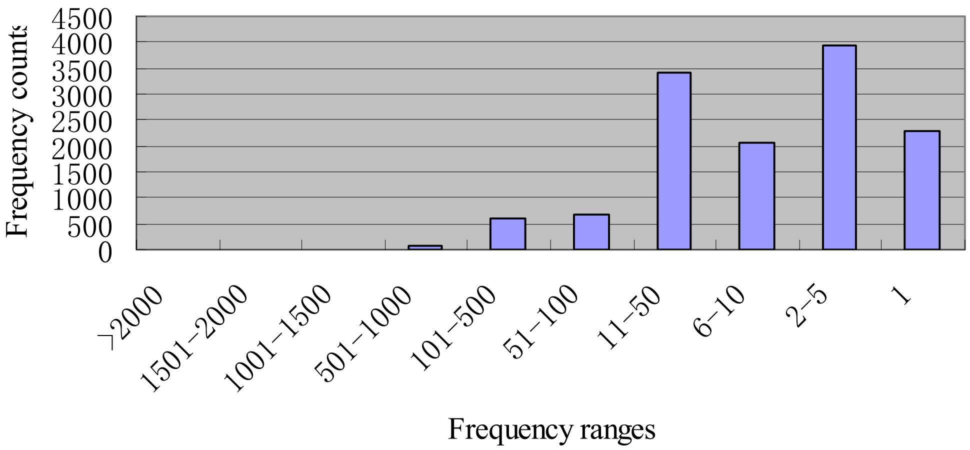
A total of 7,785,240 quadrigrams were counted, whose raw frequencies range from 1 to 4,681. The list of quadrigrams is checked against Da's consolidated list of words and phrases and 13,093 are identified as present in the list. Table 7 lists quadrigram frequency distribution among different frequency ranges, including the number of words or phrases that are also present in the HSK vocabulary list.

Table 7 Quadrigram frequency distribution

Frequency range	Raw frequency	In Da's consolidated list	In the HSK list
>2000	14	8	0
1501-2000	17	7	0
1001-1500	56	15	2
501-1000	229	60	0
101-500	4774	603	24
51-100	9254	675	23
11-50	112568	3409	70
6-10	173480	2070	22
2-5	1630692	3956	28
1	5854156	2290	9
Total	**7785240**	**13093**	**178**

In Table 7, we find that the distribution patterns of raw quadrigrams are similar to those of bigrams and trigrams, i.e., high frequency quadrigrams makes up a small portion and the majority of quadrigrams are in the very low frequency range. The distribution of those quadrigrams that are also present in Da's consolidated list has a similar distribution pattern as in the case of trigrams, i.e., they distribute more or less evenly across the medium to low frequency ranges, as shown in Figure 3.

Figure 3 Quadrigrams also present in Da's consolidated list



Random sampling of those quadrigrams that are also present in Da's consolidated list suggests that there may be a dominant number of quadrigrams are either compounds or phrases with a 2+2 structure. In addition, there seems to be quite a number of idiomatic expressions, especially in those low frequency ranges. The exact number of those two observations is yet to be determined.

3. Discussions

3.1 Comparison with Da's (2004) character list

In this study, we identified 63,64 unique characters from the news corpus. It is interesting to compare our current list with the Modern Chinese character list reported in Da (2004). Da's (2004) Modern Chinese list is based on a more than 193 million character corpus of both informative and imaginative texts. Comparing the two lists, we find that the number of unique characters used in journalistic Chinese is about two-thirds of the characters reported in Da's (2004) Modern Chinese character list. Tables 8 and 9 list the frequency distribution patterns of the two character frequency lists.

Table 8 Cumulative frequency in terms of percentages

Corpus	**10%**	**25%**	**50%**	**75%**	**90%**	**95%**	**98%**	**99%**	**99.5%**	**100%**
News corpus	7	38	155	419	838	1204	1742	2184	2651	6364
Modern Chinese	6	33	152	481	1056	1566	2284	2838	3423	9933

Table 9 Cumulative frequency in terms of individual characters

Corpus	**100**	**500**	**1000**	**1500**	**2000**	**2500**	**3000**	**3500**	**5000**
News corpus	40.6%	79.2%	92.3%	97.0%	98.7%	99.4%	99.7%	99.8%	100%
Modern Chinese	41.8%	75.8%	89.1%	94.6%	97.1%	98.5%	99.2%	99.5%	99.9%

From the two tables, we can conclude that reading current news requires less number of characters than unlimited reading in Modern Chinese. For instance, if we take 95% character coverage rate as the minimum for adequate reading comprehension, around 1,200 characters will be needed for news reading. whereas for unlimited reading in Modern Chinese it will take more than 1,500 characters. Further, the number of characters required to cover an additional 3% of texts is 500 for news texts but 700 for Modern Chinese in general.

3.2 Comparison with the HSK vocabulary list

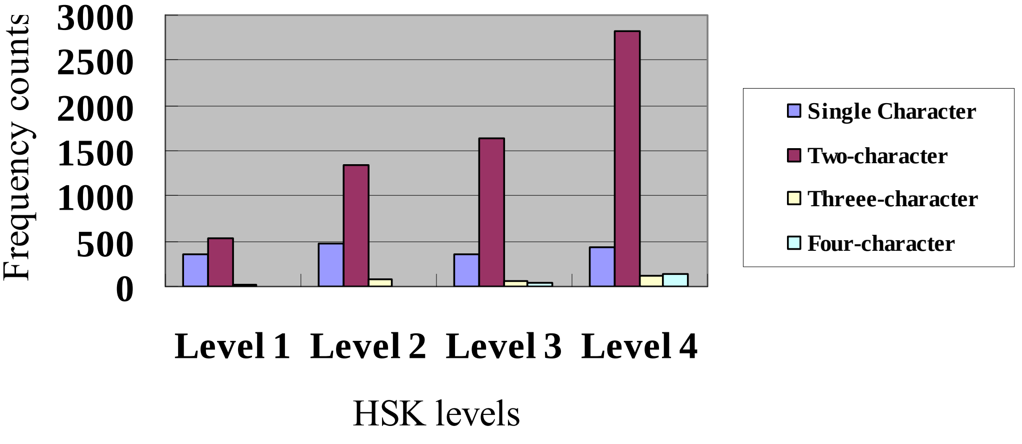
In Tables 5, 6 and 7, we presented information concerning the number of n-grams generated in this study that are also present in the HSK vocabulary list, where it can be observed that the majority of those n-grams are located in the medium to low frequency ranges. Table 10 list the distribution of those words and phrases at different HSK levels.

Table 10 HSK characters, words, phrases and idioms found in the news corpus

Characters/words/phrases	Level 1	Level 2	Level 3	Level 4	Subtotal	HSK itself
Single Character	361	464	355	434	1614	1866
Two-character	528	1343	1635	2821	6327	6373
Three-character	20	73	69	118	280	306
Four-character	2	5	30	141	178	188
Total	911	1885	2089	3514	8399	8733

In Table 10 we observe that 1) characters distribute more or less evenly across the four HSK levels; 2) nearly all disyllabic, trisyllabic and quadrisyllabic words in the HSK vocabulary list are found to be present in our news corpus; and 3) the majority of those multisyllabic words or phrases are at high HSL levels. This pattern can be seen clearly in Figure 4 below.

Figure 4 Distribution of HSK characters and words at different HSK levels



The above observations have at least two implications for CFL learning and instruction. First, it can be suggested that a learner needs to know almost all multi-syllabic words and phrases provided in the HSK vocabulary list before adequate news reading comprehension can be achieved. Given the small amount of vocabulary prescribed in the HSK list and the huge number of words and phrases found in our news corpus, it can be predicted that even if CFL learners acquired all words and phrases in the HSK vocabulary list, it would still be difficult for them to have adequate news reading comprehension. Secondly, if we are going to rely on the HSK vocabulary list as guideline for developing CFL learners' reading proficiency in Chinese news reading, it is better to begin a news reading program after a CFL learner passes HSK Level 2. Otherwise, the requirement on vocabulary knowledge may be too demanding.

3.3 How much vocabulary a CFL learner should know

In studies on collocation (in English), Mutual Information has been used as a measure of strength of association between the two elements in a bigram (Church and Mercer 1993). Before conducting this study, we had planned to use Mutual Information values to help us identify the major (if not all) of words and phrases from the list of bigrams and trigrams found in our news corpus. However, the wide range of distributions of the Mutual Information values of those bigrams found our news corpus that are also present in both the HSK vocabulary list and Da's consolidated list (as shown in Table 10 and Figures 5 and 6) indicates that there is no clear cut-off Mutual Information value so that it can be used to separate disyllabic words and phrases from those senseless bigram combinations.

Table 11 Distribution of Mutual Information values of those bigrams that are also present in the HSK vocabulary list and Da's consolidated list

Bigram lists	Range	Mean	Standard deviation
The HSK vocabulary list	-4.87 to 20.00	5.63	3.56
Da's consolidated list	-7.46 to 21.09	3.21	3.81

Note: Both the mean and standard deviation values for the two lists are based on those bigrams whose frequencies are between 6 and 30,0057 inclusive.

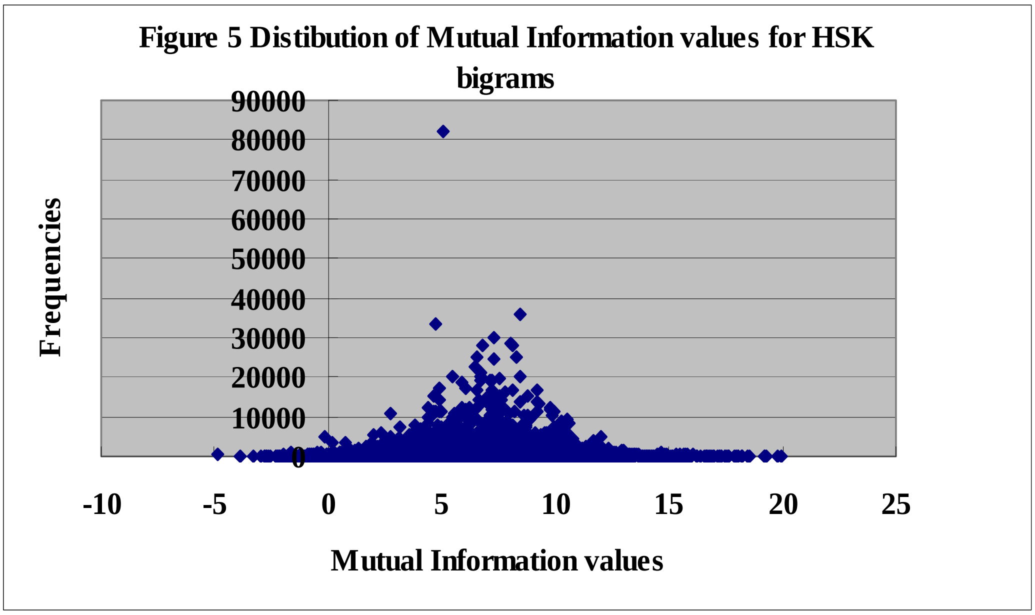
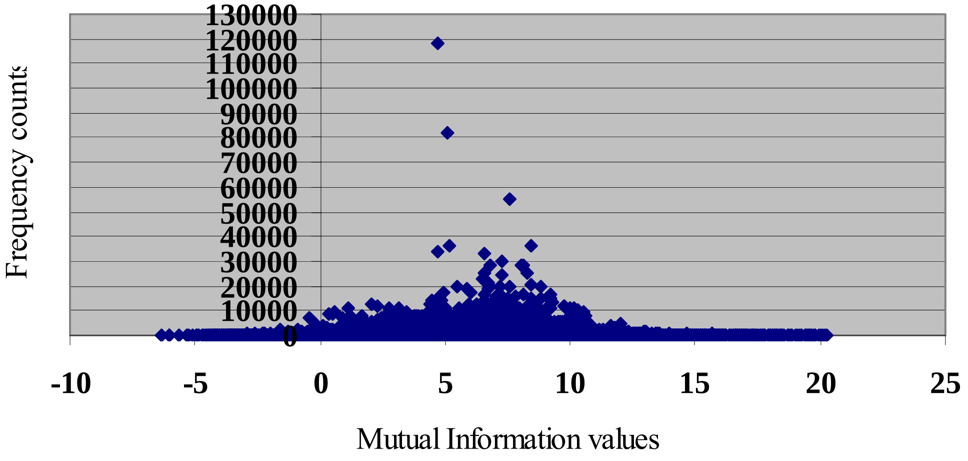


Figure 5 Distibution of Mutual Information values for HSK bigrams

Figure 6 Distribution of Mutual Information values for bigrams in Da's consolidated list



In our introduction to this paper, we mentioned that medium-sized Chinese dictionaries contain between 50,000 and 60,000 entries. In our study, we compiled a word and phrase list based on six manually edited word/phrase lists in the public domain that contains more than 140 thousand entries. While some of the entries in the list may be part of a word or phrase, we believe that given the size of our consolidated list and the size of medium-sized Modern Chinese dictionaries, those entries in our consolidated list should cover a very high percentage of words and phrases used in Modern Chinese.

In our study, 62,065 bigrams, 17,301 trigrams and 13,093 quadrigrams from news corpus are found to be present in Da's consolidated list of words and phrases. This suggests the upper limit of (multisyllabic) vocabulary used in journalistic Chinese is most likely in the neighborhood of 90,000. Given our observation that there is a tendency that many quadrigrams are compounds with a 2+2 structure, we estimate that the lower limit of vocabulary used in journalistic Chinese should be around 80,000.

While the accuracy of our estimates remains to be determined with further detailed examination of our data, the information we have obtained so far nevertheless have some implications for CFL reading instruction. On the one hand, the large amount of vocabulary used in journalistic Chinese suggests that it is impractical to teach CFL learners so many words or phrases within the time frame of any formal CFL instruction program, which in turn suggests that sufficient vocabulary acquisition for adequate news reading comprehension can only be achieved through extensive reading outside the classroom. On the other hand, if news reading instruction is to be integrated into a CFL program, it is better to offer it at the advanced level. From a pedagogical perspective, we suggest that CFL learners should aim at three targets of vocabulary acquisition if they want to develop proficiency in news reading (c.f., Table 12 and Figures 7 and 8). First, it is essential for them to acquire around 20,000 high to medium frequency (X>50) words and phrases used in journalistic Chinese. The focus of learning at this stage is on disyllabic words and phrases. Secondly, it is beneficial that they know or are exposed to around 37,000 medium-low to low frequency words and phrases in addition to the first 20,000 or so vocabulary if they want to develop a better understanding of news texts in Chinese. While the focus of learning is still on disyllabic words and phrases at this stage, more attention should be shifted onto trisyllabic and quadrisyllabic words and phrases. Thirdly, if they want full proficiency in news reading, they should be exposed to an additional 34,000 words and phrases that are rarely used in journalistic Chinese. At this stage, they should pay more or less equal attention between disyllabic and trisyllabic/quadrisyllabic words and phrases.

Table 12 Three targets for CFL vocabulary acquisition

Targets	Frequency range	2-character	3-character	4-character	Total
First	High to medium (X>50)	16545	2633	1368	20546
Second	Medium-low to low (50≥X>5)	25157	6587	5479	37223
Third	Very low (≤5)	20363	8081	6246	34690
Total		**62065**	**17301**	**13093**	**92459**

Figure 7 Three targets of vocabulary Acquisition in terms of total number of words and phrases

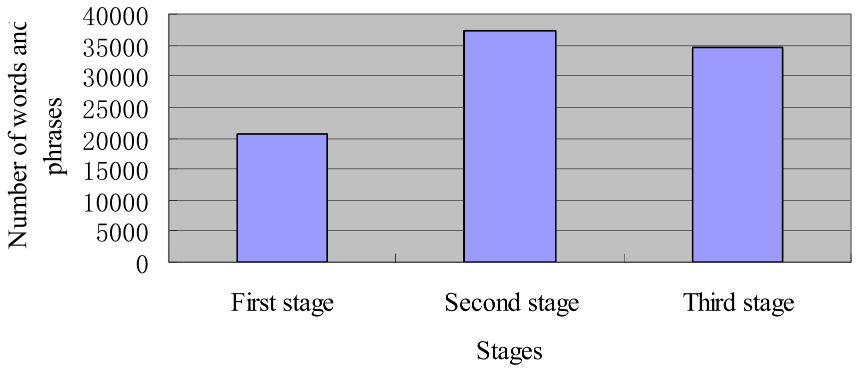
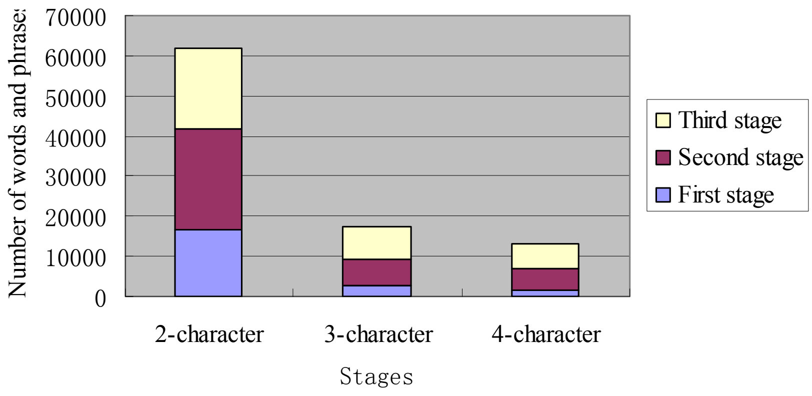


Figure 8 Three targets of CFL vocabulary acquisition in terms of the number of characters in words and phrases



4. Concluding remarks

In this paper, we reported the findings of a corpus-based study on the vocabulary used in journalistic Chinese. We compiled one character and three n-gram frequency lists based on a 20-million character corpus of news texts from various Chinese media sources in the world. With reference to a consolidated vocabulary list based on six manually edited word and phrase lists in the public domain, we estimate that the number of two- to four-character words and phrases used in journalistic Chinese is between 80,000 and 90,000. Based on our findings, we suggested that CFL learners should first focus on those high to medium frequency words and phrases (that make up about one fifth of the vocabulary identified from the corpus) if they want to develop basic proficiency in Chinese news reading. They should acquire a large amount of low frequency words and phrases if they want to further develop their news reading abilities.

It is hoped that results from this study will be useful for future reading research and instruction in Chinese. With our estimate on the vocabulary size in journalistic Chinese, for example, it remains to be discovered if Chinese native speakers do recognize those words and phrases. Information about the vocabulary used in Chinese news media (such as its size and distribution information) obtained in this study can be used as a measure based on which vocabulary tests can be designed to discover native speakers' working vocabulary size in journalistic Chinese. In terms of CFL, information about the vocabulary in journalistic Chinese can be used to help develop news reading instruction and learning materials. The same information can also be used for studies on CFL learners' deciphering process and difficulties in news reading.

As far as the current study is concerned, there are several improvements to be made. For one thing, raw texts in the corpus should be manually edited to clean up non-Simplified Chinese texts so that character segmentation can be more accurate. In addition, because of time constraints and limited human resources, no detailed examination of the n-grams is performed. Instead, our discussions and conclusions in this paper have been based on random sampling of those n-grams generated. While such an approach is acceptable in terms of statistical analysis, we believe that accurate results can only be achieved through detailed manual analysis on the data set. A third improvement is a further examination on the subsets of our news corpus. Our current research is based on an overall examination of the news texts that cover a wide range of topics in Chinese news reporting from various sources and regions. It would be interesting if we could look further into our news corpus by dividing the texts into different regions and subject areas. Such an examination on subsets of news texts would help us answer questions such as the number of words or phrases needed for adequate reading comprehension in domestic or international news or in subject areas such as politics, finance and sports, etc.

Those improvements as well as related relation questions concerning vocabulary and reading in Chinese and CFL learners will be the subjects of future research.

References

Brew, Chris and Marc Moens. 2000. Data-Intensive Linguistics. (Online version at: <http://www.ltg.ed.ac.uk/~chrisbr/dilbook/>, Last checked: 2004-03-26)

Carver, R.P. 1994. Percentage of unknown vocabulary words in text as a function of the relative difficulty of the text: Implications for instruction. Journal of Reading Behavior. 26,4,413-437

Church, Kenneth W. and Robert L. Mercer. 1993. Introduction to the special issue on computational linguistics using large corpora. Computational Linguistics. 19.1.1-24

Da, Jun. 2004. A corpus-based study of character and bigram frequencies in Chinese e-texts and its implications for Chinese language instruction. The studies on the theory and methodology of the digitized Chinese teaching to foreigners: Proceedings of the 4th International Conference on New Technologies in Teaching and Learning Chinese, ed. by Zhang, Pu, Tianwei Xie and Juan Xu, 501-511. Beijing: The Tsinghua University Press

Goulden, R., P. Nation and J. Read. 1990. How large can a receptive vocabulary be? Applied Linguistics 11: 341-363.

Hirsh, D. and Nation P. 1992. What vocabulary size is needed to read unsimplified texts for pleasure? Reading in a Foreign Language. 8,2,689-696

Hu, M. and I.S.P. Nation. 2000. Unknown vocabulary density and reading comprehension. Reading in a foreign language. 13.1. 403-430. (Online version at: <http://nflrc.hawaii.edu/rfl/PastIssues/rfl131hsuehchao.pdf>, Last checked: 2005-06-30)

Laflamme, J. G. 1997. The effect of multiple exposure vocabulary method and the target reading/writing strategy on test scores. Journal of Adolescent and Adult Literacy. 40,5,372-381

Laufer, B. 1989. What percentage of text-lexis is essential for comprehension? In C. Lauren and M. Nordamn (eds.): Special Language: From Humans Thanking to Thinking Machines. 316-323. Clevedon: Multilingual Matters.

Laufer, B. 1992. How much lexis is necessary for reading comprehension? In H. Bejoint and P. Arnaud (eds.). Vocabulary and Applied Linguistics. 126-132

Liu, Na and I.S.P. Nation. 1985. Factors affecting guessing vocabulary in context. RELC Journal. 16,1,33-42

Nagy, W. E. and J. Scott. 2000. Vocabulary processes. Handbook of reading research Vol III. (Kamil, M., etc. eds.). 269-284. Mahwah, N.J.: Lawrence Erlbaum Associates

Nation, Paul. 2001. Learning vocabulary in another language. Cambridge, UK: Cambridge University Press

Nation, Paul and Robert Waring. 1997. Vocabulary size, text coverage and word lists. In Schmitt, N. and M. McCarthy (Eds.): Vocabulary: Description, Acquisition and Pedagogy: Cambridge, Cambridge University Press. 6-19

Waring, Rob and Paul Nation. 2004. Second Language Reading and incidental vocabulary learning. Angles on the English Speaking World. 4,97-110. (Online version at: <http://www1.harenet.ne.jp/~waring/papers/waring%20120304.pdf> Last checked: 2005-05-31)

Zechmeister, E. B., Chronis, A. M., Cull, W. L., D'Anna, C. A., and Healy, N. A. 1995. Growth of a functionally important lexicon. Journal of Reading Behavior. 27,2,201-212.

冯志伟. 2002. 中国语料库研究的历史与现状. Journal of Chinese Language and Computing, 11(2) 127- 136. (Online version at <http://www.china-language.gov.cn/doc/FengZhiwei01.doc> Last checked: 2005-06-22)

汉语大词典编辑委员会. 1990. 汉语大词典(第一版). 上海：汉语大词典出版社

王惠. 汉语词汇统计研究. (Date of publication unknown from the online source. (Online version at <http://www.huayuqiao.org/articles/wanghui/wanghui06.doc> Last checked: 2005-06-20.)

中国社会科学院语言研究所词典编辑室. 1983. 现代汉语词典(第二版). 北京：商务印书馆