(Appeared in) 2004. The studies on the theory and methodology of the digitalized Chinese teaching to foreigners: Proceedings of the Fourth International Conference on New Technologies in Teaching and Learning Chinese. Zhang, Pu, Tianwei Xie and Juan Xu. (eds.). 501-511. Beijing: Tsinghua University Press.

A corpus-based study of character and bigram frequencies in Chinese e-texts and its implications for Chinese language instruction[1]

Jun Da
Middle Tennessee State University
Murfreesboro, Tennessee, USA 37132
jda@mtsu.edu

Abstract: This paper describes the findings of a research project whose main objective is to compile a character frequency list based on a very large collection of Chinese texts collected from various online sources. As compared with several previous studies on Chinese character frequencies, this project uses a much larger corpus that not only covers more subject fields but also contains a better proportion of informative versus imaginative Modern Chinese texts. In addition, this project also computes two bigram frequency lists that can be used for compiling a list of most frequently used two-character words in Chinese.

Keywords: Chinese text corpus, character, bigram, frequency, word segmentation, Mutual Information

1. Introduction

Character and word frequencies are useful information for Chinese language learning and instruction. Chinese learners are often curious about how many characters they should learn in order to master the language. Answers to the question vary from 1,000 to 3,500 or even more characters, depending on whether Chinese is learned as first or second/foreign language. Similar interests are also found among authors of Chinese language learning materials. From time to time they rely on frequency information to decide on which particular sets of characters and words to include and how to sequence them in the learning materials they develop.

In the past, character frequency information has been made available from several sources. One important source is the List of Frequently Used Characters in Modern Chinese (**《现代汉语常用字表》**, henceforth *Changyong Zibiao*) recommended jointly by the National Working Committee on Languages and Writing Systems (**国家语言文字工作委员会**) and the Ministry of Education, China in 1988. It includes 3,500 characters divided into two frequency levels. According to the Ministry of Education[2], the list was compiled based on information from other previously compiled

[1] Research for this paper was supported in part by Middle Tennessee State University Faculty Research and Creative Activity Grant Program in 2001.
[2] c.f., http://www.moe.edu.cn/moe-dept/yuxin/index.htm.

character frequency lists, dictionaries as well as a corpus of Chinese texts published from 1928 to 1986 covering ten subject categories.

In addition to the government-sponsored character list compiled in the late 1980s, there were several empirical studies on Chinese character frequency in the late 1990s whose results are accessible on the Internet. For example, Tsai (1996) compiled a character frequency list based on 1993–1994 Big5-encoded newsgroup archives. Da (1998) computed character frequency lists based on a 45 million character corpus of Simplified Chinese texts collected from various online sources. He (1998) produced a character frequency list from a trans-regional diachronic survey of Chinese literary texts published in the 1960s, 1980s and 1990s.

Apart from the above three empirical studies whose results are accessible online, there have been several other corpus-based studies of Chinese texts conducted at Peking University and Tsinghua University, etc. (Feng 2002). While those studies are reported to have looked into character and/or word frequency information one way or the other, details of their results have unfortunately not been made public and hence are beyond the reach of most Chinese language instruction professionals and researchers.

As far as the four accessible frequency lists are concerned, there are a few problems that may have hindered their usefulness. In the case of *Changyong Zibiao*, it is not known how those 3500 are ranked among themselves. If one wants to use the list for developing beginning-level learning materials, for example, it is difficult to decide which basic set of characters to include and how to present them sequentially so that learners are provided with maximum exposure to the language within the limited time frame of a language learning program. In the case of the three empirical studies mentioned above, they tend to be based on collections of Chinese texts that are either too limited in subject domains or encoded in outdated Chinese encoding standard. For example, Da's (1998) study used materials that were encoded in GB2312-80, a character set that contains less than 7,000 distinct characters. While they provided useful information at the time of the study, it now appears that the Da's (1998) results are outdated given the fact the more and more Chinese webpages are encoded in the more recent GB13000 (also known as GBK) or GB18030[3] standards which contain much larger character sets.

As compared with the availability of detailed information about character frequencies, information about word frequencies is much scarce and only a few graded word lists are accessible to Chinese language learners and instruction professionals. One such graded list is the HSK List of 8,000 Chinese Vocabulary published by Beijing Language and Culture University in 2000. The other is Dew's (1999) handbook which grades 6,000 Chinese vocabulary into elementary, intermediate and more advanced levels. The scarcity of accessible information about word frequencies in Chinese may be due to the fact that while a word in Chinese may contain one, two, three or even more characters, researches employing heuristic methods for segmenting individual words in running Chinese texts that do not contain word delimiters are far from conclusive (c.f., Sproat and Emerson 2003, among others).

---

[3] For more information about various encoding standards for Chinese characters, please refer to, for example, http://www.praxagora.com/lunde/cjk_inf.html.

In this paper, we report the findings of a recent research project whose main objective is to compile yet another character frequency list based on a very large collection of online Chinese texts that are encoded in not only the GB2312-80 but also the GB13000 standard. With detailed results of the research project made available at `http://lingua.mtsu.edu/chinese-computing`, we will focus our discussion in this paper on the construction of the corpus used in the study and some general distribution patterns of the character frequencies found in our corpus. In addition, we will discuss the computing of two bigram frequency lists that can be used as the basis for compiling a two-character word frequency list in Modern Chinese. It is hoped that those frequency lists will provide a better tool for both Chinese language learners and instruction professionals.

2. This study

2.1. Corpus design and data collection

The main objective of this research project is to compile a character frequency list that can be used for both Chinese language learning and instruction. Accordingly, the following three measures have been taken in the construction of the Chinese text corpus used in the study: 1) Both Classical and Modern Chinese are collected from various online sources, where texts written before 1911 are categorized as Classical and those published in or after 1911 Modern Chinese. 2) Only formal Modern Chinese texts are included in the corpus. No efforts have been made to collect informal writings of Modern Chinese such as postings on various online BBS or email messages. 3) With references to the structures of Brown Corpus (Francis and Kucera 1964), British National Corpus (Burnard 2000) and Longman/Lancaster Corpus (Summers 1991), efforts are made to collect text materials from a diverse range of subject fields (c.f. Table 1). In addition, a distinction is made between imaginative (i.e., those written for entertainment or related to literary works) and informative texts (i.e., those written for information and/or knowledge) for Modern Chinese.

Table 1: List of subject fields used in the study

Category	Subcategory	Subject fields
Classical Chinese		Novels, prose, history, poetry and drama, etc.
Modern Chinese	Informative	Computer science, economics, education, government, health, history, law, military, news, philosophy, politics, popular science, religion, etc.
	Imaginative	General fiction, children, detective, drama, history, Kongfu or martial arts, military, prose, literary review and science fiction, etc.

All electronic texts used in this study were collected between 1997 and 2003 from various online sources (c.f., Table 2) that fall into two different categories. On the one hand, some websites such as **中国青少年新世纪读书网** and **国学网络**, etc. offer large collections of digitized texts that were originally published in printed format. Examples of this kind of texts include Classical and Modern Chinese texts published before 1995. On the other hand, other websites such as **中国科普博览** and **免费医院网**, etc. provide original texts that are written and published on the Internet and are

intended for online readers. Examples of this kind of texts include online newsletters and magazines and online course learning materials, etc. Given the large quantities of data used in this study, it is impractical for us to manually download and examine each individual webpage or online file. Instead, tools such as *w3mir*[4] were used to automatically harvest texts after samples of their contents were examined. In our text colletion, cautions were taken to harvest only those webpages or other kinds of online files that contain useful text data.

Table 2: Sources of Chinese electronic texts

Sources	URL
中国青少年新世纪读书网	http://www.cnread.net
亦凡公益图书馆	http://www.shuku.net
国学网络	http://www.guoxue.com
素心书斋	http://member.netease.com/~luolian/suxinshuzhai.htm
新语丝	http://www.xys.org
文学视界	http://www.white-collar.net/index.asp
国家统计局	http://www.stats.gov.cn/tjgb/index.htm
国家人口和计划生育委员会	http://www.chinapop.gov.cn
农业部中国农业信息网	http://www.agri.gov.cn
中国科普博览	http://www.kepu.com.cn
南京博物院	http://www.njmuseum.com/zh/book/wbsk.htm
高教出版社	http://wljx.hep.edu.cn
免费医院网/天天健康报	http://www.cmn.com.cn
人民日报	http://snweb.com/
计算机世界	http://www.ccw.com.cn
电脑报	http://www.cpcw.com

2.2 Data processing

All computing tasks are performed on the FreeBSD[5] platform using both Unix Shell commands and customized scripts written in the PHP scripting language[6]. MySQL[7] is used as the backend database software.

2.2.1 Data pre-processing

The majority of our text collections are HTML-encoded webpages. Before character frequencies were counted, HTML tags were automatically removed from those texts with the help of a built-in function of the PHP scripting language. After the removal of HTML tags, another customized PHP script was used to automatically remove any text strings used for website navigation

[4] c.f., http://langfeldt.net/w3mir/
[5] c.f., http://www.freebsd.org
[6] c.f., http://www.php.net
[7] c.f., http://www.mysql.com

or as webpage footers. Examples of website navigation strings include **回首页，上一页** and **下一页**, etc., whereas instances of webpage footer strings include **亦凡书库** and **站长信箱**, etc. It is reasonable for us to assume that those text strings are mostly (if not all) automatically included on webpages or online files and do not belong to the original written texts that are interest to us.

2.2.2 Character segmentation

Unlike western languages such as English, individual characters or words in running Chinese texts are not delimited with any whitespaces. Hence, the reliability of any methods for automatic character segmentation will depend on the encoding scheme found in those text files. A quick inspection of our text corpus shows that the majority (if not all) of the texts used in this study are encoded in GB2312-80 or GB13000, both of which employ a two-byte encoding scheme. Since the majority of the electronic texts in our corpus were harvested via some automatic method, it is difficult for us to tell if any of those files are encoded in the most recent GB18030 encoding standard, in which a character may be encoded using two or four bytes. Our best estimate is that even if there were some files encoded in GB18030 in our text collection, their numbers would be very small and hence will not affect significantly the frequency distribution patterns reported in this paper.

2.2.3 Bigram counting

In this study, a bigram is defined as a string with two consecutive characters in a text and can be treated as a close approximation to a two-character word in Chinese. To compute bigram frequencies, we used a modified method based on Brew and Moens (2000). All running Chinese texts were first segmented into segments of continuous character strings where both GB encoded symbols and ASCII codes were treated as delimiters. Bigrams were then identified and counted within each continuous character string.

It should be pointed out here that bigrams are a super-set of two-character words in Chinese. 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 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 (c.f., Church and Mercer 1993). Accordingly, we also computed Mutual Information scores for those bigrams found in this study. While Mutual Information values alone may not be enough to distinguish between those meaningful and non-sense two character combinations, we believe that when both raw bigram frequency and Mutual Information score are used at the same time, they will help producing a rather accurate list of most frequently used two-character words in Chinese.

2.3 Results

5 frequency lists were compiled in this study: 1) A character frequency list for Classical Chinese; 2) A character frequency list for Modern Chinese; 3) A combined character frequency list for both Classical and Modern Chinese; 4) A bigram frequency list (with Mutual Information values) based on the news sub-corpus; and 5) A bigram frequency list (with Mutual Information values)

based on the general fiction sub-corpus. Due to space considerations, details of those frequency lists as well as other relevant information about the corpus are made available at `http://lingua.mtsu.edu/chinese-computing`. In this section, we will restrict our discussions to the general distribution patterns of both character and bigram frequencies observed in the study.

2.3.1 Character frequency distribution

Table 3 lists some summary statistics of our character frequency count. As can be seen in the table, over 258 millions of characters are identified from our collection of Chinese e-texts, in which Modern Chinese makes up more than 193 million and Classical Chinese more than 65 million. Further, informative texts make up 55% of our Modern Chinese collection and imaginative 45%. This is comparable to the Longman/Lancaster Corpus which is specially designed to study the lexicon of Modern English.

Table 3: Summary statistics of character counts

Corpus	Total number of characters	Number of unique characters
Classical Chinese	65348624	11115
Modern Chinese	193504018	9933
Informative	106254415	8954
Imaginative	87249603	8435
Total	258852642	12041

Among those 258 million characters, 12041 unique or distinct characters are identified in both Classical and Modern Chinese, where Modern Chinese contains 9,933 and Classical Chinese 11,115, respectively. Appendices A and B list 1,000 most frequently used characters of Classical Chinese and Modern Chinese, respectively.

Tables 4 and 5 provide a summary of cumulative frequency information about the more than 258 million characters counted. More specifically, Table 4 lists the number of unique characters in terms of cumulative frequency percentiles, whereas Table 5 lists cumulative frequency percentiles in terms of the number of unique characters. It is interesting for us to highlight the fact here that the top 1,056 characters account for 90% of our Modern Chinese collection, even though 9,933 distinct characters are identified.

Table 4 Cumulative frequency in terms of percentages

Category	Cumulative frequency								
	10%	25%	50%	75%	90%	95%	99%	99.5%	100%
Classical	12	53	220	703	1598	2433	4432	5094	11115
Modern	6	33	152	481	1056	1566	2838	3423	9933
Combined	7	39	179	573	1264	1891	3590	4367	12041

Table 5 Cumulative frequency in terms of individual characters

Corpus	Accumulative frequency								
	100	500	1000	1500	2000	2500	3000	3500	5000
Classical	34.8%	67.7%	82.0%	89.0%	93.0%	95.3%	96.8%	97.8%	99.4%
Modern	41.8%	75.8%	89.1%	94.6%	97.1%	98.5%	99.2%	99.5%	99.9%
Combined	38.9%	72.1%	86.2%	92.4%	95.6%	97.3%	98.3%	98.9%	99.7%

2.3.2 Bigram frequency distribution

Because researches in automatic identification of Chinese words using statistical methods are still far from perfect, we only counted bigram frequencies and computed their Mutual Information values in the news and the general fiction sub-corpora, where the former contains 14 million and the latter 18 million characters. Those two sub-corpora can be considered representatives of informative and imaginative texts in Modern Chinese, respectively and are used in this study to explore the possibility of generating two-character word frequency lists with the help of Mutual Information values.

Table 6 lists summary statistics concerning bigrams found in those two sub-corpora, where we find that their raw frequencies range from 1 to more than 50,000. A breakdown of raw frequency ranges for those bigrams is listed in Table 7, where we find that the majority of bigrams have a raw frequency less than 10.

Table 6 Bigram counts in the news and general fiction sub-corpora

Corpus	**Total characters**	**Distinct characters**	**Unique bigrams**	**Total bigrams**	**Raw frequency range**	**Average frequency range**
News	14339418	5990	730067	12470872	1 - 53185	17.1
General fiction	18070786	6489	973338	15807413	1 - 57186	16.2

Table 7 Number of unique bigrams in 6 frequency ranges

Corpus	**X>=1000**	**500<= X<1000**	**100<=X< 500**	**50<=X<100**	**10<=X<50**	**X<10**
News	1506	1709	13450	16004	96175	601223
General fiction	1816	2252	18277	19788	111196	820009

As far as individual Mutual Information values are concerned, they range from 24.11 to -10.82 for the general fiction sub-corpus and 23.77 to -10.39 for the news sub-corpus. Based on Da's (1998) preliminary observation that a bigram with Mutual Information value equal or greater than 3.5 is a good word candidate, Table 8 lists the number of distinct bigrams in 6 raw frequency ranges where individual Mutual Information values are equal or greater than 3.5.

Table 8 Number of unique bigrams in 6 frequency ranges where Mutual Information >=3.5

Corpus	X>=1000	500<= X < 1000	100<=X< 500	50<=X<100	10<=X<50	X< 10
News	1104	921	4433	3628	15424	77036
General fiction	822	864	6053	5514	20707	110172

Appendices C and D list the top 1,000 bigrams for the news and general fiction sub-corpora where individual Mutual Information values are equal or greater than 3.5. An informal inspection of the complete bigram lists shows that those bigrams with raw frequencies at 50 or higher and Mutual Information values at 3.5 or higher are good candidates for two-character words in both sub-corpora.

2.4 Discussions

2.4.1 Comparison with *Changyong Zibiao*

The government-sponsored *Changyong Zibiao* lists 3,500 most frequently used characters in Modern Chinese. According to the Ministry of Education, the 3,500 characters would cover 99.48% of a 2 million character corpus, where the first 2,500 characters accounted for 97.97% of and the remaining 1,000 1.51%.

For the sake of comparison, we also looked up frequency information about those 3,500 words in our corpus whose details are listed at `http://lingua.mtsu.edu/chinese-computing`. It turns out that the first set of 2,500 characters counts a total of 188,748,366 characters, covering 97.54% of our Modern Chinese corpus. The remaining 1,000 characters count a total of 3266243 or 1.69% of our Modern Chinese corpus. In terms of individual character frequencies, we find that among those 3,500 characters, a total of 314 characters out of the two frequency sets have frequencies below 1,000 with the lowest at 17 (for the character 柒). In contrast, the 3,500th character in our Modern Chinese list has a frequency of 1,033. That there are characters with very low frequencies included in both subsets of *Changyong Zibiao* may be due to the fact that when the list was compiled, factors such as their linguistic functions and occurrences across subject domains were also considered.

2.4.2 Implications for Chinese language instruction

The character and bigram frequency lists compiled in this study have several implications for Chinese language learning and instruction. First, information from the character frequency lists can be used for specifying the learning outcomes for each stage of a language instruction program. For example, a minimal set containing the top 1,500 characters can be targeted for beginning-level instruction, a basic set containing the top 2,500 characters for intermediate-level instruction and an expanded set containing the top 3,500 to 5,000 characters for advanced-level instruction. Secondly, with detailed frequency information provided, stratified random character samples can be selected from those frequency lists for improving language tests that measure learners' true knowledge of Chinese characters. Thirdly, our bigram lists, though still require further screening, suggest that the number of frequently used two-character words in Modern Chinese most likely fall in the range of

10,000 to 12,000, whereas the number of high frequency two-character words is around 2,000. Based on further screening of individual bigrams, we can recommend three sets of two-character words containing the top 2,000, the top 6,000 and the top 10,000 for different stages of Chinese language learning.

3. Concluding remarks

In this paper, we have described the compiling of both character and bigram frequency lists based on a large corpus of Chinese electronic texts. As compared with several previous researches on Chinese character frequencies, this study offers two improvements: 1) It uses a much larger corpus that not only covers more subject fields but also contains a better proportion of informative versus imaginative texts. Hence, we believe that its results provide a better picture of Chinese character frequencies in real language use; 2) Based on bigram frequency information and their Mutual Information values obtained in this study, it is now possible for us to come up with a list of frequently used two-character words in Chinese.

There are several improvements to be made about the current study. First, better sampling method can be used to select a true representative of Chinese texts. In this study, we tried to select texts from as many subject fields and authorship as possible. However, the whole piece rather than a portion of any selected texts is included in our computing, where their lengths can range from several hundreds to more than 10 thousand characters. The varied length among those sample texts may have skewed true character and bigram frequency distributions. Secondly, while we assumed that all texts in our corpus are encoded in either the GB2321-80 or GB13000 standard, it would be beneficial to identify texts that are encoded in the GB18030 standard so that more accurate counts of characters and bigrams can be conducted. Lastly, our text corpus only includes Chinese texts written in formal style. No efforts have been made to include informal writings such as postings on BBS or email messages. Such limited choice of texts may have failed to provide a complete picture of real language use. It is hoped that future research on Chinese character frequencies can be improved along the lines suggested above.

References

北京语言文化大学汉语水平考试中心. 2000. HSK **汉语** 8000 **词词典**. **北京**: **北京语言文化大学出版社**

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: 2004-03-29)

国家语委, **教育部**. 1988. **现代汉语常用字表** (**见《语言文字规范手册》**). **北京：语文出版社**. (Online version at: <http://www.moe.edu.cn/moe-dept/yuxin/index.htm>, Last checked: 2004-03-29)

何秀煌. 1998. **香港、大陸、台灣 - 跨地區、跨年代**: **現代漢語常用字頻率統計** (Hong Kong, Mainland China & Taiwan: Chinese Character Frequency - A Trans-Regional, Diachronic

Survey). (Online version at: <http://www.arts.cuhk.edu.hk/Lexis/chifreq/>, Last checked: 2004-03-29)
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)
Burnard, Lou. 2000. The British National Corpus Users Reference Guide (Online version at: <http://www.natcorp.ox.ac.uk/World/HTML/urg.html>, Last checked: 2004-03-11)
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. 1998. Chinese text computing. (Online version at: <http://lingua.mtsu.edu/chinese-computing/old-version1/>, Last checked: 2004-03-29).
Dew, James Erwin. 1999. 6,000 Chinese Words: A Vocabulary Frequency Handbook. Boston, MA: Cheng & Tsui Company
Kennedy, Graeme. 1998. An Introduction to Corpus Linguistics. London: Addison Wesley Longman
Leech, G. 1992. Corpora and theories of linguistic performance. in Directions in corpus linguistics (Svartivik ed.). 105-122. Berlin: Mouton de Gruyter
Francis, W.N. and H. Kucera. 1964. Manual of Information to accompany A Standard Corpus of Present-Day Edited American English, for usewith Digital Computers (revised 1979). Providence, Rhode Island: Department of Linguistics: Brown University (Online version at: <http://helmer.aksis.uib.no/icame/brown/bcm.html>, Last checked: 2004-03-11)
Sproat, Richard and Thomas Emerson. 2003. The First International Chinese Word Segmentation Bakeoff. Paper presented at The Second SIGHAN Workshop on Chinese Language Processing. Sapporo, Japan. (Online version at: <http://www.sighan.org/bakeoff2003/paper.pdf>, Last checked: 2004-03-29)
Summers, D. 1991. Longman/Lancaster English Language Corpus: Criteria and Design. Harlow: Longman
Tsai, Chih-Hao. 1996. Frequency and Stroke Counts of Chinese Characters. (Online version at: <http://www.geocities.com/hao510/charfreq/>, Last checked: 2004-03-29)

Appendix A List of 1,000 most frequently used characters in Modern Chinese

Note: Characters are listed from left to right in order of descending frequency.

的一是不了在人有我他这个们中来上大为和国地到以说时要就出会可
也你对生能而子那得于着下自之年过发后作里用道行所然家种事成方
多经么去法学如都同现当没动面起看定天分还进好小部其些主样理心
她本前开但因只从想实日军者意无力它与长把机十民第公此已工使情
明性知全三又关点正业外将两高间由问很最重并物手应战向头文体政
美相见被利什二等产或新己制身果加西斯月话合回特代内信表化老给
世位次度门任常先海通教儿原东声提立及比员解水名真论处走义各入
几口认条平系气题活尔更别打女变四神总何电数安少报才结反受目太
量再感建务做接必场件计管期市直德资命山金指克许统区保至队形社
便空决治展马科司五基眼书非则听白却界达光放强即像且难权思王象
完设式色路记南品住告类求据程北边死张该交规万取拉格望觉术领共
确传师观清今切院让识候带导争运笑飞风步改收根干造言联持组每济
车亲极林服快办议往元英士证近失转夫令准布始怎呢存未远叫台单影
具罗字爱击流备兵连调深商算质团集百需价花党华城石级整府离况亚
请技际约示复病息究线似官火断精满支视消越器容照须九增研写称企
八功吗包片史委乎查轻易早曾除农找装广显吧阿李标谈吃图念六引历
首医局突专费号尽另周较注语仅考落青随选列武红响虽推势参希古众
构房半节土投某案黑维革划敌致陈律足态护七兴派孩验责营星够章音
跟志底站严巴例防族供效续施留讲型料终答紧黄绝奇察母京段依批群
项故按河米围江织害斗双境客纪采举杀攻父苏密低朝友诉止愿细千值
仍男钱破网热助倒育属坐帝限船脸职速刻乐否刚威毛状率甚独球般普
怕弹校苦创假久错承印晚兰试股拿脑预谁益阳若哪微尼继送急血惊伤
素药适波夜省初喜卫源食险待述陆习置居劳财环排福纳欢雷警获模充
负云停木游龙树疑层冷洲冲射略竟范句室异激汉村哈演策简卡罪判担
州静退既衣您宗积余痛检差富灵协角占配征修皮挥胜降阶审沉坚善妈
刘读啊超免压银买皇养伊怀执副乱抗犯追帮宣佛岁航优怪香著田铁控
税左右份穿艺背阵草脚概恶块顿敢守酒岛央托户烈洋哥索胡款靠评版
宝座释景顾弟登货互付伯慢欧换闻危忙核暗姐介坏讨丽良序升监临亮
露永呼味野架域沙掉括舰鱼杂误湾吉减编楚肯测败屋跑梦散温困剑渐
封救贵枪缺楼县尚毫移娘朋画班智亦耳恩短掌恐遗固席松秘谢鲁遇康
虑幸均销钟诗藏赶剧票损忽巨炮旧端探湖录叶春乡附吸予礼港雨呀板
庭妇归睛饭额含顺输摇招婚脱补谓督毒油疗旅泽材灭逐莫笔亡鲜词圣
择寻厂睡博勒烟授诺伦岸奥唐卖俄炸载洛健堂旁宫喝借君禁阴园谋宋
避抓荣姑孙逃牙束跳顶

Appendix B List of 1,000 most frequently used characters in Classical Chinese

Note: Characters are listed from left to right in order of descending frequency.

之不一人以有了为道是子的来大也十其上二而中曰下于三得在年我他
王说事见将者去日天州出后又自此时个无军太这与月所家如知你里公
行可使到四至五那官书生小言何相兵今都南山就只正前明东着门帝等
文国百西心安然要好入回同从老两进已马起云元当及未臣便还地夫城
过看令儿多六用方士乃分长金平能故则命万头先史名皆主死司成问武
外复水八请七师宗高千诸听内若朝身尚李和作北定县张女因阳面数府
江即德亦把河本守礼九世君众白开闻走立并意部亲才处皇常必法非会
欲们手民谓么声且少遂路诏玉叫位氏郎左甚余风弟度初义各杀父归侍
花口神发第光取坐应遣矣学京间卷春酒字重没几右祖罢既经往再汉宝
黄边通笑御节宫却对兴周母真石陈话打尽气直情传海齐难乐己拜吾清
合夜做居刘龙郡奏钱置功封住原兄想虽爷举实岁向贼卿被敢领议望思
别制注许送信些姐次政受求秋加古衣改服娘远侯解林流给理孙卒忠奉
首留房战连喜罪称物新善刺食宜教广更章户吃或观报赐保建引力关色
殿随降圣任宋治台告乱半旧计待放青省乎始吏带宣转病姓香杨失妇银
动论宁卫孝但由甲差镇吴魏诗总交寻足陵怎骑属管迁早客仁久监深贤
员尔贵星莫景语车堂赵收秦代唐集飞破仙反近修异梁备丁曾营志果焉
院空夏华火全急临样田忽参永岂终止易每忙恩率辞利点谢容除历致晋
拿败征伯胜妻承庙仪指接凡兼阴怀表倒遇徐满执密红庆授须耳恐济幸
土犯游督惊船丞召害散免福攻木尉讨性化号贾愿申良俱觉乘廷舍干抚
落候美按轻顺胡徒记犹献什谁依退谋职雨玄目独帅郑步姑尝提罗竟爱
素辛绝眼延变比强品汝英灵校毕追答辰设谷惟威哥迎击室昌益极亡怒
权伏仍敬推商骸救升副图疑约察僧虎恶细登业梦统刑似固阿运视凤布
淮音精寿楼乡楚牛达草略尊离据乙朔苦列共疾午群断川曹持赏昭嘉陛
奇叔射招特隐旨顾肯克句疏饮快奔盖湖曲私体晚祭康巡势暗伤禁怕呢
纳苏欢仆珠结越休呼刻丑念庚写工孔微燕让庄剑般友彼斩选息婆怪寺
韩仲壬哭双录禄妹照崇吉严希著盛象条厚丹丙隆养哉丈废宰期闲刀考
阁鼓戊富宿纪造算敌试歌器鲁铁蒙围雄鬼决调掌班驾奴爵馀务秀柳虚
识丧贞兰诛荣沙寅累完树端典癸盗俊获巳阵博诚末假系形载继冬扬显
寇场唬斗雪施悉感赠劳酉卯讲番移藏遗劝赤具嗣朕陆谏忧种奸存晓缘
饭策男补机充护黑紫案恭索现冠泰画贡萧屯术亥芳勿乞佛杜洪单袭赶
减积咸件妖薄岳药谕切伐叹刚状允园惠戊争朱松办尺旁邑寒读颜吕荆
印钟赴范律狱静邪防庶睡雷恨认词投赞况财野逃旗拔惧席辅市孟茶奶
怨冲屋脱背买费宾孤洛渐党判夷亮童附负乌梅役辽粮托俗撰赋贺勇量
祀穿迹笔敕检幽跟脚专

Appendix C List of 1,000 most frequently used bigrams in news sub-corpus

Note: Bigrams are listed from left to right in order of descending frequency where Mutual Information values are equal or greater than 3.5.

中国 美国 发展 经济 国家 问题 一个 工作 台湾 社会 政府 记者 我们 人民 进行 北京
企业 表示 国际 他们 没有 建设 关系 代表 世界 公司 全国 报道 市场 日本 组织 技术
重要 方面 合作 地区 领导 活动 拉克 认为 教育 日电 伊拉 管理 目前 已经 自己 有关
会议 总统 研究 安全 人员 委员 政治 可以 要求 通过 可能 中央 加强 一些 计划 部门
主义 本报 政策 改革 今年 联合 同时 就是 第一 指出 生产 科技 文化 干部 这些 提出
大陆 成为 这个 今天 举行 提高 主要 情况 支持 历史 群众 投资 解决 两国 一步 发生
现在 主席 香港 服务 科学 学生 时间 新华 开始 实现 希望 建立 以及 工程 大学 如果
公投 积极 信息 开发 作为 增长 农民 消息 环境 必须 因为 布什 提供 水扁 陈水 保护
外交 影响 农业 国务 新闻 华社 美军 和平 军事 生活 需要 决定 基础 思想 产品 调查
精神 民主 中心 到了 实施 法律 系统 世纪 访问 继续 会主 不能 战略 总理 这种 民族
包括 产业 罗斯 贸易 部分 事件 上海 其中 这样 部长 参加 项目 什么 地方 图片 机构
一次 特别 去年 制度 俄罗 城市 农村 以来 年来 国内 但是 基本 规定 正在 共同 美元
增加 两岸 资源 报告 利益 稳定 官员 开展 由于 努力 不断 媒体 第二 专家 员会 学习
行动 监督 武器 出现 负责 坚持 目标 也是 务院 统一 接受 取得 能力 二十 调整 台独
责任 双方 战争 导弹 得到 因此 强调 作用 中共 法轮 委会 其他 行政 发现 反对 严重
根据 机关 许多 轮功 水平 单位 措施 国防 人士 重点 了解 采取 结构 现代 印度 己的
使用 任何 处理 促进 而且 三个 英国 全面 重大 十五 成功 经营 领域 应该 收入 非常
进入 基地 利用 他说 部队 这次 犯罪 对于 亿元 期间 之间 方式 任务 实际 法院 不同
具有 一种 朝鲜 万元 出口 受到 新网 全球 推动 工业 发言 造成 关于 发表 不仅 达到
电视 存在 还是 集团 事业 规模 开放 完成 银行 最大 恐怖 一直 斯坦 条件 家宝 温家
中华 达姆 书记 只有 萨达 都是 比赛 形成 分子 竞争 两个 对台 导人 江泽 所有 发挥
亚洲 人才 还有 第三 泽民 航天 创新 介绍 意见 会见 维护 旅游 资金 一名 结果 获得
友好 作出 内容 当地 过去 一位 依法 不可 起来 主任 参与 推进 如何 同志 原则 保障
来自 重视 也不 公安 消费 飞机 执行 成员 韩国 之后 宣布 西部 甚至 充分 先进 十分
过程 经过 保持 虽然 能够 会谈 质量 根本 准备 公开 学校 情报 袭击 时候 各种 正式
看到 分析 常委 未来 力量 建议 改变 成立 扩大 传统 亿美 人口 所以 当局 并不 是否
体制 有效 完全 别是 标准 独立 综合 加快 最后 交流 知识 原因 直接 明年 结束 透露
分别 实行 最近 网络 日前 机制 关注 深入 按照 长期 邪教 电话 控制 打击 个月 卫星
制定 显示 不过 文明 结合 财政 明确 金融 健康 规范 欧洲 据新 宣传 医院 阶段 最高
当时 高度 相关 威胁 欧盟 青年 召开 制造 形势 色列 言人 以色 另一 广大 针对 防部
解放 体系 优势 事实 立法 对此 全部 检察 发射 立场 创造 人权 很多 意义 具体 劳动
爆炸 万人 落实 完善 认真 下午 保证 济发 自治 交部 良好 基金 首次 帮助 此次 各地
价格 检查 运动 这里 采访 报记 比较 政协 办公 知道 卫生 变化 选举 体育 理论 公里
时代 持续 生态 日消 经贸 设施 代化 次会 超过 产生 只是 欢迎 讨论 设计 系列 凤凰
附图 随着 新世 认识 多云 形式 面对 文章 各级 个代 题上 十二 协会 军队 家庭 带来

实践 时期 引起 举办 协议 一项 海军 经验 或者 民党 人类 卫视 考虑 更多 更加 出席
副主 自然 关键 案件 自由 表明 及其 百分 明显 规划 共产 关部 凰卫 才能 意识 真正
艺术 场经 祖国 事务 共和 选择 电子 素质 出版 不少 无法 改善 投入 联系 上午 华盛
孩子 居民 盛顿 表现 告诉 民币 小时 利亚 汽车 确定 斗争 地位 西藏 组成 部署 执法
成果 现场 公布 广泛 对外 石油 办法 强烈 年代 集中 论坛 增强 突出 培训 职工 业务
现象 安排 死亡 势力 就业 社北 感到 巨大 十年 儿童 最终 产党 来说 表团 应当 基层
电影 减少 空间 飞行 分之 态度 反映 生命 来越 革命 左右 相信 越来 每年 在伊 困难
进口 指导 有限 再次 说明 价值 统计 大量 报报 纽约 近日 队伍 京市 首先 近年 训练
发布 当前 火箭 面临 新社 警方 专业 学院 大使 范围 昨天 除了 一切 尽管 秩序 相当
拥有 责人 手段 报讯 进步 司法 小组 整个 另外 一样 如此 程度 仍然 奥运 一条 航空
伊朗 很大 资料 导致 交通 装备 符合 做好 治理 进党 事情 然而 广东 需求 先生 利于
道德 贡献 正确 只要 力度 大常 攻击 指挥 资产 加入 此外 第四 多次 多数 所谓 讲话
切实 土地 一批 今后 三十 严格 确保 挑战 妇女 先后 相互 高级 将于 台海 粮食 设备
因素 批准 机场 那么 那些 社记 锦涛 评论 胡锦 做出 展开 协调 声明 销售 坚决 内部
突破 涉及 医疗 非法 人数 巴格 状况 保险 湾问 承诺 民众 几个 留学 台北 报导 法规
体现 阿拉 事故 主权 移民 副总 几年 数字 批评 商业 格达 特色 令人 迅速 签署 商品
合理 大规 贯彻 各项 星期 方案 法制 民群 同意 区域 矛盾 逐步 权利 其它 必要 上述
进程 学者 培养 少年 武装 动力 尤其 改造 注意 破坏 战斗 危机 方向 之前 运会 首都
给予 教授 分裂 压力 正常 广州 作战 转变 这位 其实 面积 朋友 秘书 宪法 军方 九届
联盟 效益 关心 公室 义务 反应 革开 药品 建筑 计算 违法 报北 用于 限制 研制 一般
局势 成绩 础上 日讯 清楚 士兵 乡镇 大型 海外 当然 太空 十六 恢复 座谈 适应 现实
前往 贫困 治安 形象 重庆 曾经 年轻 这项 广告 纪念 交易 科研 党委 修改 据悉 西方
立即 现状 铁路 观念 况下 道路 五十 集体 教科 镕基 际上 导干 及时 朱镕 基斯 始终
放军 局长 业化 购买 冠军 对话 方针 邀请 运输 愿意 化建 证明 呼吁 南京 资本 内地
贷款 承认 危险 非洲 演习 警察 网站 社区 速度 大选 团结 回答 附近 巴基 四川 残疾
只能 警告 新技 下降 华民 李鹏 南省 从事 合同 密切 专门 核心 国驻 还要 拿大 主张
数据 谈判 党员 五年 紧张 费者 权力 理解 浙江 育活 拒绝 加拿 不足 难以 防御 空军
不久 等方 丰富 普遍 尊重 职业 潜艇 申请

Appendix D List of 1,000 most frequently used bigrams in general fiction sub-corpus

Note: Bigrams are listed from left to right in order of descending frequency where Mutual Information values are equal or greater than 3.5.

一个 什么 没有 自己 我们 他们 知道 起来 这个 时候 这样 怎么 已经 现在 出来 不能
还是 不知 可以 女人 觉得 因为 你们 孩子 那个 一点 这么 可是 一种 两个 眼睛 心里
下来 那么 还有 东西 先生 这些 看见 一次 也没 父亲 地方 回来 生活 这种 过去 人家
母亲 一切 男人 工作 只有 声音 今天 开始 时间 几个 如果 中国 但是 以后 有什 有些
事情 所以 突然 那些 儿子 一定 为什 问题 有点 许多 太太 告诉 感到 电话 想到 朋友
坐在 于是 大家 老师 并不 下去 说话 看到 当然 家里 明白 可能 第一 一直 而且 然后
书记 应该 身上 似乎 那样 看看 喜欢 出去 脸上 最后 还没 妈妈 晚上 意思 一条 别人
姑娘 虽然 一句 学生 望着 世界 以为 好像 也许 发现 站在 给你 会儿 一块 多少 样子
感觉 女儿 第二 一阵 面前 正在 之后 爸爸 得很 二十 其实 想起 希望 一片 咱们 一般
非常 总是 完全 公司 然而 学校 忽然 个女 回去 年轻 走到 日子 所有 老板 怎样 仿佛
十分 回到 不敢 让我 关系 只要 高兴 笑着 自然 回答 听见 回家 句话 几乎 女孩 终于
准备 主任 小姐 听到 衣服 甚至 需要 让他 离开 一件 情况 清楚 门口 领导 相信 同学
那种 任何 办法 手里 带着 马上 另一 口气 实在 身体 几天 同时 愿意 进来 当时 都没
发生 日本 里面 精神 丈夫 原来 哪里 对于 个月 接着 社会 路上 三个 如何 或者 如此
件事 特别 房子 认为 干部 认识 出现 只能 大学 到底 北京 轻轻 它们 办公 就像 不愿
床上 老太 仍然 慢慢 同志 上海 之间 吃饭 头发 立刻 国人 刚才 老婆 跟着 容易 本来
汽车 无法 永远 并且 找到 进去 点儿 痛苦 不管 身边 结婚 十年 站起 医院 外面 除了
老头 明天 妻子 因此 走进 变成 走出 放在 青年 医生 这位 方面 名字 根本 目光 注意
队长 只好 很多 每天 房间 显得 故事 生命 以前 点头 屋里 时代 渐渐 很快 一步 太阳
决定 常常 继续 走过 媳妇 肯定 爱情 眼泪 鬼子 哪儿 留下 机会 能够 眼前 并没 革命
结果 三十 变得 后面 发出 父母 公室 无论 感情 思想 旁边 敌人 成为 真正 打开 多年
土司 简直 微笑 这件 几年 房里 亦铭 心中 经过 经理 如今 记得 前面 个男 奶奶 重要
下午 漂亮 大概 城市 坐下 难道 要求 郑亦 直到 脑袋 答应 中间 整个 意识 文化 干什
尽管 二天 事儿 定要 幸福 一段 爷爷 在床 哥哥 不错 街上 人物 肚子 必须 可怜 作为
见过 沉默 充满 姐姐 来越 越来 桌上 么办 至于 进行 新月 文学 再也 人民 历史 之中
早就 曾经 作者 家庭 露出 生意 中学 丽萍 小孩 客人 教授 阳光 表示 谁也 好象 立即
去找 奇怪 少平 半天 院子 嘴里 夫人 心情 才能 躺在 国家 屋子 点点 局长 全部 切都
辈子 哈哈 鲜花 摇头 耳朵 美国 重新 两只 等着 消息 紧张 嘴唇 不懂 年前 忘了 个字
说完 确实 老爷 由于 妹妹 天晚 悄悄 分钟 鼻子 台上 主义 师傅 手指 块钱 桌子 不久
回头 参加 校长 好几 解释 双手 另外 穿着 男孩 部分 实际 研究 大哥 笑道 韩梅 村里
少安 理解 家伙 当年 早已 即使 玻璃 公羊 昨天 部队 究竟 星期 不肯 随着 火车 电影
之前 脸色 兴奋 五十 认真 司令 政治 艺术 毕业 想象 放心 拿出 对方 去吧 诉我 抱着
坐着 十几 激动 回事 存在 用手 海鹏 同意 东方 介绍 简单 意见 刚刚 态度 既然 其中
城里 毫无 把它 果然 责任 美丽 不断 一股 十二 一副 报告 放下 作家 只手 不必 原因

十多 拿着 紧紧 以及 志杰 门外 小伙 电视 不仅 令人 习惯 各种 尤其 政府 力量 田志
晓得 艾雨 害怕 失去 兄弟 起身 男子 出门 室里 农民 抬起 夜里 第三 差不 每个 将来
考虑 脚步 生气 忘记 主要 热闹 跑到 底下 事实 转过 英明 转身 朱海 安排 十五 从前
情绪 依然 作品 活动 胳膊 那位 快乐 脖子 四十 表情 面对 反正 一遍 比较 招呼 显然
红色 一番 一层 几句 对面 最好 接受 组织 平静 打算 自由 睡觉 等待 表现 知识 结束
魏强 静地 抓住 范英 黑暗 三天 问道 消失 几次 一座 默默 掏出 担心 石头 几十 一辈
教育 剩下 比如 老洪 帮助 劳动 进入 经常 居然 封信 伸出 不停 妇女 命运 随便 王国
伙子 商量 弟弟 警察 行动 忍不 聪明 指挥 年纪 打电 背后 脑子 内心 少年 大夫 休息
周围 司机 华丽 学习 诉你 三年 请你 意义 受到 报纸 男女 更加 提出 见面 影响 盯着
交给 百惠 发展 林百 主意 此刻 似地 所谓 心理 平时 天空 微微 慢地 眼光 产生 不及
怀疑 赶紧 宿舍 谈话 开口 国炎 干净 百姓 使她 单位 道理 解决 沙发 怀里 此时 仔细
命令 公社 地区 泪水 熟悉 身后 不够 界上 代表 一辆 院里 到处 向前 灯光 变化 现实
薄荷 农村 照片 椅子 李丽 空气 竟然 车站 田福 少爷 难以 同样 收拾 属于 些什 恋爱
人类 恐怕 屁股 欣然 早晨 满意 两年 指着 安慰 毕竟 战争 一顿 外国 关于 浑身 窗外
白色 土地 是否 一颗 到哪 困难 轻地 至少 糊涂 陌生 成功 理由 就算 承认 文章 今年
一份 负责 传来 白天 一旦 往往 谢谢 老汉 病人 深深 老先 努力 可怕 远远 故意 热情
通过 想法 音乐 匆匆 从此 运动 地站 摇摇 位置 沉重 离婚 读书 喝酒 经济 伸手 点钟
神情 选择 演习 政委 大约 拉着 机关 厉害 懂得 脾气 笑容 接过 开门 等等 上班 厨房
得更 计划 兴趣 拿起 左右 静静 人群 视着 经验 感动 改变 很难 颜色 直接 首先 灵魂
主席 玩笑 郑浩 联系 送到 检查 情形 毫不 年龄 中午 拒绝 远处 很少 老实 麻烦 却又
坏了 战斗 科长 无数 客气 冈普 绝对 十岁 可惜 际上 识到 事业 众人 唯一 要紧 一套
睡着 乡下 任务 仅仅 教师 遇到 跟前 方向 陈老 维民 方式 抬头 县长 不幸 个世 肩膀
寻找 公安 英雄 中央 危险 墙上 互相 夜晚 弄得 罗维 亲自 模样 每次 条件 十八 楼上
监狱 十万 年代 定会 证明 经历 门前 绣文 如同 笑笑 一杯 动作 大队 群众 袋里 最近
几分 叶子 巨大 处长 太多 庄稼 宗三 公路 孙少 生产 子奇 反而 满足 回忆 卜绣 印象
技术 能让 犹豫 秘书 相当 影子 行为 会议 加上 普通 半夜 张老 一封 默地 夫妻 飞机
必要 某种 无所 环境 队员 引起 工资 马路 严重 韩子 迅速 秘密 记者 赶快 他俩 十六
处理 空中 关心 往下 伤心 强烈 乡长 狠狠 两条 严肃 月亮 眼镜 婆婆 完成 意外 舒服
安静 房门 子寒 上午 时刻 口袋 洪子 能力