FREQUENTLY-USED CHINESE CHARACTERS AND LANGUAGE COGNITION

Chin-Chuan Cheng
University of Illinois at Urbana-Champaign
City University of Hong Kong
c-cheng1@uiuc.edu

To facilitate the spread of literacy in China, several lists of frequently-used Chinese characters have been published in the last 70 years. The number of characters in each list does not exceed 8,000. The number of character types in each Chinese book such as the 25 Dynasty Histories published in the past 2,000 years also does not exceed this number. Even when synchronous dictionaries already accumulated over 30,000 distinct characters, writers of large volumes normally used only 4,000 to 8,000 characters. The range of 4,000 to 8,000 morphemes then is proposed as the optimal number of linguistic symbols for human manipulation.

1. Number of frequently-used characters

Comprehensive Chinese dictionaries list over 50,000 distinctive Chinese characters. Many of these characters are no longer in use. From the articles in J. Wang 1995 I collected the following counts of frequently-used characters in practical lexicons or pedagogically oriented frequency books. (These references are cited in J. Wang 1995 and are therefore not given in our own References.) These counts range from 2,000 to slightly over 7,000. The titles of the volumes are given to help the reader understand their orientation.

(1)  4,261  Chen Heqin 陈鹤琴 (1928, Practical Lexicon for Colloquial Style 語体文應用字彙)
2,000  Sichuan Education College 四川省教育科學院 (1946, Frequently Used Characters 常用字選)
3,000  Education Department of Shantong Province 山東省教育廳 (1958, List of Frequently-Used Characters in Putonghua 普通話常用字表)
3,100  Secondary and Primary School Textbook Editorial Board of Education Bureau of Beijing City 北京市教育局中小學教材編審處 (1965, List of Frequently Used Characters 常用字表)
4,444  Chinese Character Section of Committee on Chinese Lan-
The term ‘frequently-used characters’ may give an impression that the number of characters was drastically reduced for elementary learning and literacy promotion. However, T’sou et al. 1997 show that current Chinese newspapers in Taipei, Hong Kong, and Singapore used about 4,000 distinct characters. It is therefore reasonable to say that a few thousand characters will suffice for linguistic expressions of most matters and events in modern times.

When we examine books of earlier times, we still see a similar number of characters used. For example, the novel Dream of the Red Mansion (紅樓夢) from the 18th century used only slightly over 4,000 characters. As it is quite certain that the first 80 and the last 40 chapters were written by two different individuals, I used the electronic version offered by Yuanze University of Taiwan to tabulate the characters separately as follows:

<table>
<thead>
<tr>
<th>(2) CHARACTER</th>
<th>CHARACTER</th>
</tr>
</thead>
<tbody>
<tr>
<td>TOKENS</td>
<td>TYPES</td>
</tr>
<tr>
<td>496,855</td>
<td>4,293</td>
</tr>
<tr>
<td>234,980</td>
<td>3,217</td>
</tr>
<tr>
<td>731,835</td>
<td>4,501</td>
</tr>
</tbody>
</table>

The entire book had over 730,000 character tokens. But only 4,501 distinct characters were used. In the same 18th century, the dictionary Kangxi Zidian (康熙字典) published in 1716 already collected a total of 47,035 distinct characters. The fact that out of the available 47,000 distinct characters only a few thousand of them were employed in a novel full of poems and descriptions of social institutions and personal feelings should be regarded as something significant about human use of linguistic symbols. In the following sections I will show that during the past 2,000 years the number of distinct Chinese characters increased from 9,000 to over 56,000. But individual authors used only a few thousand characters. This small number has been a constant across historical stages.

2. Accumulation of Chinese characters

The oracle writing of 3,300 years ago had about 5,000 distinct characters. It is generally agreed that 3,000 of them are now recognizable. Over time, additional characters were created, and the 1986 edition of the Hanyu Da Zidian (漢語大字典) had over 56,000 entries. The linear increase of distinct characters
can be clearly seen in the following list of dictionaries. The publication dates and the number of characters covered were gathered from Hsieh et al. 1992 and Yin & Rohrenow 1994:

(3) DATE CHARACTERS
1300BC 3,000 Oracle Writing 甲骨文
5AD 5,340 Xuncuanpian 訓纂篇
100 9,353 Shuowen Jiezi 説文解字
4C. 12,824 Zilin 字林
543 16,917 Yupian 玉篇
1008 26,000 Guangyun 廣韻
1039 31,000 Leipian 類篇
1067 30,000 Jiyun 集韻
1615 33,179 Zihui 字匯
1670 33,000 Zhengzhitong 正字通
1716 47,035 Kangxi Zidian 康熙字典
1915 48,000 Zhonghua Da Zidian 中華大字典
1959 49,965 Da Han-He Cidian 大漢和辭典
1968 49,905 Zhongwen Da Cidian 中文大辭典
1986 56,000 Hanyu Da Zidian 漢語大字典

The sudden jump within a short span, for example, from a total of 48,000 characters in the year 1915 to 49,965 in 1959 and from 49,905 in 1968 to 56,000 in 1986 could not be attributed to the additional characters introduced during the intervals. The differences have to do with the comprehensiveness of the coverage of the dictionaries. Some editors had more resources to collect a larger number of characters. However, over the past 2,000 years, the gradual increase was obvious in this listing.

As users of dictionaries, we often feel that many of the characters or words in them are totally unfamiliar. Unabridged dictionaries are supposed to collect all words, frequent as well as rare, for users to look up. So it is natural for us to encounter unfamiliar characters. As we know, all aspects of language change in the course of history. Dictionary compilers are scholars who work on words unfamiliar to them; they have to look up words in other dictionaries in the compilation process. Why do we have unfamiliar words? Over a century ago, Darwin attributed the reason to a limit to the powers of the memory: ‘We see variability in every tongue, and new words are continually cropping up; but as there is a limit to the powers of the memory, single words, like whole languages, gradually become extinct’ (Darwin 1871:94-5). We will come back to this point later.

Not all words or characters in a dictionary are unknown to us. We may be able to derive the senses of some characters in contexts, but we have never used them before. We may say that these characters are in our passive vocabulary. There are familiar ones, of course. They are in our active vocabulary. How many are there of familiar ones? That is, how many characters do we know? This question has to do with human language cognition. Earlier we saw that 4,000 may be considered as an optimal number for language manipulation. We will pursue this
issue further in the following sections.

3. How many words do you know?

Traditionally a way to find the answer to the question ‘how many words do you know?’ is through experiments. In such experiments subjects were asked to identify known words from randomly selected pages of a dictionary, and the statistics were projected to cover the entire repertoire of the lexicon (Crystal 1995). Crystal 1995 states that English speakers can have 31,500 to 56,250 words in their active vocabulary and 38,300 to 76,350 words in their passive vocabulary. Miller & Gildea 1991 state that in the United States high school graduates at age 17 normally have 80,000 words in their vocabulary. Thus in the first 16 years of life they acquired 5,000 words per year or 13 words per day on average.

Do we indeed have 31,000 or 80,000 words in our active vocabulary? Are we actually able to use that many words? I have proposed to use the number of words in various books to help find the answer (Cheng 1997, 1998). First of all, in English the inflectional endings change a word into several graphic forms. For example, the nine words below are derived from the basic forms of ‘write’ and ‘kick’:

(4) write, writes, wrote, writing, written, kick, kicks, kicked, kicking

We can say that these nine words are ‘graphic words’ and the two basic forms are the ‘concept words’. I wrote a computer program to lemmatize words by returning words ending in -s, -ed, -ing, -ly, -er, -est and various forms of pronouns to the basic forms. Following is a list of words that show the inflected and basic forms:

(5) ABILITIES: ABILITY
     ABRASIVELY: ABRASIVE
     ABRASIVES: ABRASIVE
     BUILDERS: BUILDER
     BUILDING: BUILD
     BUILDS: BUILD
     BUILT: BUILD
     COOLED: COOL
     COOLER: COOL
     COOLEST: COOL
     COOLING: COOL
     ME: I

I ran the program on scores of electronic English texts obtained from the Web. The numbers show a consistent pattern. No matter how long the books are, the concept words used range from about 4,000 to 8,000 in number. Some examples are given below:

(6) TOKEN | GRAPHIC | CONCEPT | TITLE
         |         |         | Call of the Wild
32,361  | 4,727   | 3,431   | Tom Sawyer
74,038  | 7,427   | 5,316   |
C. C. CHENG: FREQUENTLY-USED CHINESE CHARACTERS

87,044  8,630  6,046  Beauty and the Beast
161,751  9,281  6,433  Dracula
137,060  8,877  6,218  The American
39,631  4,236  3,190  Aspens Papers
80,493  8,976  6,377  Paradise Lost
161,974  7,097  4,647  Emma
120,735  6,288  4,199  Sense and Sensibility
123,270  6,288  4,146  Pride and Prejudice
84,128  5,741  3,934  Persuasion
729,792  13,765  8,641  Austen's 6 Books

The last set of numbers is for Jane Austen's six books combined. The books are Emma, Sense and Sensibility, Pride and Prejudice, Persuasion, Mansfield Park, Northanger Abbey. The novels by Austen covered different areas of human activities, and yet the total number of distinct words used is just over 8,000.

In Chinese, we can also examine individual authors' use of characters to determine the number of morphemes that we can actively control. Earlier we showed that the Chinese novel Dream of the Red Mansion used only 4,501 characters. Here we will examine the 25 Dynasty Histories. These Histories recorded all the activities of the emperors, ministers, and local officials, conditions of economy and society, and others. Generally speaking they are not skewed to one particular aspect of human life. The numbers of characters used in the 25 Dynasty Histories are taken from the tabulation by Hsieh et al. 1992. The publication dates, character tokens, and character types are given below. We see that the number of characters used rarely exceeds 8,000:

<table>
<thead>
<tr>
<th>DATE</th>
<th>TOKENS</th>
<th>TYPES</th>
</tr>
</thead>
<tbody>
<tr>
<td>93BC</td>
<td>533,505</td>
<td>5,122</td>
</tr>
<tr>
<td>83AD</td>
<td>742,298</td>
<td>5,833</td>
</tr>
<tr>
<td>289</td>
<td>377,807</td>
<td>4,388</td>
</tr>
<tr>
<td>445</td>
<td>894,020</td>
<td>6,161</td>
</tr>
<tr>
<td>488</td>
<td>811,893</td>
<td>5,842</td>
</tr>
<tr>
<td>514</td>
<td>299,257</td>
<td>4,962</td>
</tr>
<tr>
<td>554</td>
<td>998,329</td>
<td>5,417</td>
</tr>
<tr>
<td>636</td>
<td>294,438</td>
<td>4,973</td>
</tr>
<tr>
<td>636</td>
<td>163,382</td>
<td>4,033</td>
</tr>
<tr>
<td>636</td>
<td>212,506</td>
<td>4,032</td>
</tr>
<tr>
<td>636</td>
<td>262,659</td>
<td>4,161</td>
</tr>
<tr>
<td>648</td>
<td>1,158,126</td>
<td>5,997</td>
</tr>
<tr>
<td>656</td>
<td>701,698</td>
<td>5,592</td>
</tr>
<tr>
<td>659</td>
<td>677,624</td>
<td>5,376</td>
</tr>
<tr>
<td>659</td>
<td>1,106,543</td>
<td>5,572</td>
</tr>
<tr>
<td>945</td>
<td>2,002,600</td>
<td>6,346</td>
</tr>
<tr>
<td>974</td>
<td>790,879</td>
<td>5,109</td>
</tr>
<tr>
<td>1060</td>
<td>1,694,794</td>
<td>6,771</td>
</tr>
<tr>
<td>1072</td>
<td>291,476</td>
<td>3,909</td>
</tr>
<tr>
<td>1344</td>
<td>296,254</td>
<td>4,071</td>
</tr>
</tbody>
</table>

(7)
<table>
<thead>
<tr>
<th>Year</th>
<th>Characters</th>
<th>afüral</th>
<th>Language</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>1344</td>
<td>931,070</td>
<td>5,264</td>
<td>Jin Shi</td>
<td>金史</td>
</tr>
<tr>
<td>1345</td>
<td>3,980,123</td>
<td>7,389</td>
<td>Song Shi</td>
<td>宋史</td>
</tr>
<tr>
<td>1370</td>
<td>1,611,849</td>
<td>5,854</td>
<td>Yuan Shi</td>
<td>元史</td>
</tr>
<tr>
<td>1739</td>
<td>2,802,544</td>
<td>7,124</td>
<td>Ming Shi</td>
<td>明史</td>
</tr>
<tr>
<td>1927</td>
<td>4,514,567</td>
<td>8,080</td>
<td>Qing Shi Gao</td>
<td>清史稿</td>
</tr>
</tbody>
</table>

The token column shows the length of each book. In the past 20 centuries, even though some of the books could be as short as 163,000 or as long as 4,514,000 characters, the total number of distinct characters used was between 4,000 and 8,000. This constant range existed when the dictionaries had a total number of character types of 5,000 in the 1st century, 12,000 in the 4th century, 16,000 in the 6th century, 30,000 in the 11th century, 33,000 in the 17th century, 47,000 in the 18th century, and 56,000 in the 20th century. The entries in the lists in (3) and (7) are now combined and arranged in chronological order to show the increase in the number of characters collected in dictionaries and the constant number of characters used in Dynasty Histories over time:

(8) -1300 3,000 Oracle Writing 甲骨文
-93 5,122 Shi Ji 史記
5 5,340 Xuncuanpian 訓纂篇
83 5,833 Han Shu 漢書
100 9,353 Shuowen Jiezi 說文解字
289 4,388 Sanguo Zhi 三國志
4C 12,824 Zilin 字林
445 6,161 Houhan Shu 後漢書
488 5,842 Song Shu 宋書
514 4,962 Nanqi Shu 南齊書
543 16,917 Yupian 玉篇
554 5,417 Wei Shu 魏書
636 4,973 Liang Shu 梁書
636 4,033 Chen Shu 陳書
636 4,032 Beiqi Shu 北齊書
636 4,161 Zhou Shu 周書
648 5,997 Jin Shu 晉書
656 5,592 Sui Shu 隋書
659 5,376 Nan Shi 南史
659 5,572 Bei Shi 北史
945 6,346 Jiu Tang Shu 舊唐書
974 5,109 Jiu Wudai Shi 舊五代史
1008 26,000 Guangyun 廣韻
1039 31,000 Leipian 類篇
1060 6,771 Xin Tang Shu 新唐書
1067 30,000 Jiyun 集韻
1072 3,909 Xin Wudai Shi 新五代史
1344 4,071 Liao Shi 遼史
1344 5,264 Jin Shi 金史
1345 7,389 Song Shu 宋史
1370  5,854  Yuan Shi 元史
1615  33,179  Zihui 字汇
1670  33,000  Zhengzitong 正字通
1716  47,035  Kangxi Zidian 康熙字典
1739  7,124  Ming Shi 明史
1915  48,000  Zhonghua Da Zidian 中华大字典
1927  8,080  Qing Shi Gao 清史稿
1959  49,965  Da Han-He Cidian 大汉和辞典
1968  49,905  Zhongwen Da Cidian 中文大辞典
1986  56,000  Hanyu Da Zidian 漢語大字典

The numbers were plotted in Figure 1 (see Appendix) to show the small range of variation in the number of characters used in the Dynasty Histories in contrast with the large numbers of characters in dictionaries. While dictionary compilers collected known as well as unfamiliar words in a volume, authors of books used familiar words to express ideas or to describe events. The small range of variation in the number of characters used in the Chinese writings and in the number of words used by English writers compels us to conclude that the optimal number of linguistic symbols a person can handle is between 4,000 and 8,000.

I have made a cross-language comparison of English and Chinese and concluded that the optimal number of linguistic symbols for manipulation is between 4,000 and 8,000. However, English ‘words’ and Chinese ‘characters’ are not identical linguistic units. We will discuss this issue below.

4. Characters, and words, and linguistic symbols

As is well known, a Chinese character represents a morpheme, and a morpheme can be analyzed as bound or free. A free morpheme is a word. A large number of Chinese characters represent words. But Chinese words may consist of more than one morpheme and therefore more than one character in writing. The English words that we tabulated above are lemmatized basic forms. They may be single morphemes. Some may consist of stems and affixes.

In terms of internal structure some English words may not be identical to Chinese characters. We now examine occurrences of ‘words’.

Francis & Kucera 1982 show that the Brown University’s English corpus built in the 1960s had over 30,000 distinct basic words. However, they give only 5,996 words with a running frequency of five or above in their listing and consider the others rare words. It is interesting to note here that this number 5,996 is within our optimal range.

There are numerous words in Chinese. ‘Cidian’ (word dictionaries), in contrast to ‘Zidian’ (character dictionaries), normally give a separate count of monosyllabic (single character) and polysyllabic (multiple character) entries. Some of the following ‘word’ dictionaries have more than 90,000 distinct entries:
<table>
<thead>
<tr>
<th>(9)</th>
<th>MONO-SYLLABIC</th>
<th>POLY-SYLLABIC</th>
<th>TITLE</th>
</tr>
</thead>
<tbody>
<tr>
<td>100,000</td>
<td>10,000</td>
<td>30,000</td>
<td>Chinese Comprehensive Dictionary Editorial Board 中國大辭典編纂處 (1937, Guoyu Cidian 國語辭典)</td>
</tr>
<tr>
<td>10,000</td>
<td>6,000</td>
<td>50,000</td>
<td>He 何容 (1976, Guoyu Ribao Zidian 國語日報字典)</td>
</tr>
<tr>
<td>56,000</td>
<td>3,994</td>
<td>90,000</td>
<td>Beijing Foreign Language College English Department 北京外國語學院 英語系 (1978, Chinese-English Dictionary 漢英詞典)</td>
</tr>
<tr>
<td>2,116</td>
<td>4,000</td>
<td>58,000</td>
<td>Chinese Academy of Social Science Linguistics Institute 中國社會科學院 語言所 (1980, Modern Chinese Dictionary 現代漢語詞典)</td>
</tr>
<tr>
<td>4,000</td>
<td>9,700</td>
<td>48,000</td>
<td>Fu and Chen 傅興嶺、陳章煥 (1982, Frequently-Used Word Dictionary 常用詞語字典)</td>
</tr>
<tr>
<td>13,000</td>
<td>80,000</td>
<td></td>
<td>Liu 劉源 (1984, Modern Chinese Word List 現代漢語詞表)</td>
</tr>
<tr>
<td>6,000</td>
<td>3,994</td>
<td>90,000</td>
<td>Zhang 張聿忠 (1986, Words 詞林)</td>
</tr>
<tr>
<td>800</td>
<td>2,116</td>
<td>90,000</td>
<td>Li et al. 李國炎等 (1988, New Chinese Dictionary 新編漢語詞典)</td>
</tr>
<tr>
<td>13,000</td>
<td>36,000</td>
<td></td>
<td>Wu 吳昌恒 (1988, Old and Modern Chinese Practical Dictionary 古今漢語實用詞典)</td>
</tr>
<tr>
<td>80,400</td>
<td>800</td>
<td>80,000</td>
<td>Chinese Pinyin Lexicon Editing Group 漢語拼音詞匯編寫組 (1991, Chinese Pinyin Lexicon 漢語拼音詞匯)</td>
</tr>
<tr>
<td>13,000</td>
<td>11,000</td>
<td>28,000</td>
<td>Beijing Foreign Language University English Department 北京外國語大學 英語系 (1995, Chinese-English Dictionary 漢英詞典)</td>
</tr>
<tr>
<td></td>
<td>Ye and Huang 葉立群、黃成德 (1996, Longman New Advanced Chinese Dictionary 郎文中文高級新辭典)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Min 閔龍華 (1997, Modern Chinese Usage Dictionary 現代漢語用法詞典)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Many of the words here are combinations of morphemes and other words. They do not have to be memorized. Moreover, the commonly used ones are not as numerous as the figures would show. For example, T’sou et al. 1997 found over 40,000 ‘words’ in the newspapers in Singapore, Hong Kong, and Taipei. But the highest occurring 5,000 to 7,000 words cover 90% of the texts:
Thus we can say that even in terms of words, the optimal number is not very large. We will have to study the morphological organization of words in Chinese in detail to be able to differentiate those that have to be memorized and those that can be derived from morphological rules and patterns.

5. Language cognition

We have used ‘characters’, ‘graphic words’, and ‘concept words’ in printed matters to argue for the view that the human capacity for manipulation of linguistic symbols has a limit around 8,000 units. We see no problem in extending this view to the capacity of those who are not literate. As we know, with effort and specialization, some people may have a larger vocabulary. But literate or not, mature native speakers of any language possess similar powers of language use.

Earlier I cited Darwin’s view that ‘as there is a limit to the powers of the memory, single words, like whole languages, gradually become extinct’. The phrase ‘powers of the memory’ is perhaps the right word. We use a finite set of elements to make up other words. The words are combined to make up phrases and sentences. The powers for combination are fairly high. But I have shown that there is a limit to morphemes that we can actively manipulate. This limit cannot be due to the finiteness of physical memory locations or cells for the reason that bilinguals or multilinguals generally have the same powers for each language. If physical memory locations were the determining factor, then those who speak another language would have to replace the memory of what had been occupied by the first language. Therefore the limit of powers is a matter of memory function rather than memory location.

New words appear, and old words become extinct. This is a fact of language. I have proposed a level of optimal manipulation of linguistic symbols for the maintenance of a constant 4,000-8,000 range of characters in the 25 Dynasty Histories over time. In daily life, names of acquaintances fade away with time. With effort we can retrieve them. So are words. Various linguistic functions and activities associated with the proposed optimal number of symbols for manipulation can be profitably studied in conjunction with the studies of memory and other aspects of cognition.

REFERENCES


——. 1998. Learning words with many texts. The Proceedings of the First Inter-


Figure 1. Number of Character Types Used in Dictionaries and Twenty-Five Dynasty Histories