AUTOMATIC COMPILATION OF MODERN CHINESE CONCORDANCES

Syunsuke UEMURA*, Yasuo SUGAWARA*
Mantaro J. HASHIMOTO**, Akihiro FURUYA***

*Electrotechnical Laboratory, 1-1-4 Umezono, Sakura, Ibaraki 305, JAPAN
**Tokyo University of Foreign Studies, 4-51-21 Nishigahara, Kita, Tokyo 114, JAPAN
***Tokyo Metropolitan University, 1-1-1 Yakumo, Meguro, Tokyo 152, JAPAN

An automatic indexing experiment in Chinese is described. The first very large volume of modern Chinese concordances (two sets of one million-line KWIC index) has been compiled and materialized automatically with a modified kanji printer for Japanese.

INTRODUCTION

This paper describes an experiment to compile Chinese concordances automatically. A very large volume of KWIC indexes for modern Chinese (one million lines per set) has been compiled successfully with a kanji printer for Japanese. This paper discusses the purposes of the experiment, selection and input of the Chinese data, some statistics on Chinese characters (vs. kanji) and the concordance compilation process. Finally, examples from the computer-generated concordances are shown.

THE PURPOSES

The idea of machine-processing modern Chinese data originally came from Professor Yuen Ren Chao, Agassiz Professor Emeritus of Oriental Languages at the University of California at Berkeley, before one of the authors (Hashimoto) took over the directorship of the Princeton Chinese linguistics project. Chao served as the chief of the advisory committee to the project since its foundation. The idea, in short, was: so much has been said about the Chinese pai-hua-wen -- a written language of modern China -- yet nobody has ever clarified what it really was, i.e.; what the basic vocabulary was, what the major syntactic structure was, etc.: in other words the every detail of the reality of pai-hua-wen. Certain quantitative surveys were done before us, but even the most extensive one in those days was based on data consisting of no more than 100,000 characters. In addition, the selection was very poorly done -- most of the materials were primary school textbooks. We did not believe that school textbooks reflected the reality of the language, even in its written form. We chose one digit more than the previous one, namely 1,000,000 characters, though for various reasons, the actual data contained in our tape include several thousands more than one million [1, 2].

After completion of the computer input and editing of the million-character file at Princeton, researches towards statistical aspects of the data have been conducted [4]. As stated in [4], tables of character frequency can tell us various aspects of the Chinese, such as the basic character set, transient states of character strings and so on. This can be summarized as the first step of computer-processing modern Chinese data. However, in order to understand the reality of a language, besides statistics, concordances are the necessities which illustrate the contexts where and how those characters are used.

On the other hand, computer applications to Chinese have very limited background so for. No computer-generated concordances on Chinese have been reported yet. Thus the concordance genaration project would not only be valuable to the understanding of Chinese pai-hua-wen, but also contribute to the development of the methodology to manipulate Chinese automatically. Consequently, a project to compile concordances of the Princeton million-character file was conducted at the Electrotechnical Laboratory during 1977-1979. This constitutes the second important stage of computer-processing modern Chinese.

THE CHINESE DATA

The Input of the Original Data

The first phase of the data input was done in Taiwan during 1969-1972 with a Chinese character keyboard, designed by Cheng Chin Kao -- a Chinese teletype Machine (manufactured by the Oki Denki Co., Ltd.). The code was converted into the Chinese standard telegraphic code in Walthum, Massachusetts at a computer company. The greatest difficulty, in addition to ordinary proofreading, consisted in the conversion of the so-called "combination characters" of the C.C.Kao system: any character not found in the Kao keyboard was punched so that part of it (normally the "radical") was represented by a character having the same radical in the keyboard, and another by a character having the same "signific". Necessary flags were of course attached to these "combination characters", yet the key punchers selected those constituent characters quite at random, sometimes

disregarding the position of a radical within a character, so that the results were often a hopeless mess.

The Selection of the Data

It was tried, at the selection of the data, to cover every conceivable category and style of writings in China since her modernization, the so-called May 5 Movement period, from ordinary novels to philosophical writings, from political speeches to newspaper articles, etc. etc. These categories and styles were classified and were assigned appropriate marks to show the genre. The partial list of these writings follow:

魯迅: 阿Q正伝	端木蕻良: 大江	陳独秀: 今日中国之政治問題
巴金: 家	張愛玲: 金鎖記	政務院: 中央人民政府政務院命令
茅盾: 子夜	趙樹理: 李家莊的変遷	溥儀: 我的前半生
老舍: 駱駝祥子	郭沫若: 金剛坡下	毛沢東: 実践論
艾蕪: 石青嫂子	梅蘭芳: 舞台生活四十年	林彪: 人民戦争勝利万歳
曹禺: 北京人	孫逸仙: 三民主義	劉少奇: 関于土地改革問題的報告
銭鍾書: 囲城	李済: 殷墟建築遺存報告序	周恩来: 関于知識分子問題的報告
沈従文: 辺城	胡風: 文芸工作的発展及其努力方向	聞介石: 行的道理

For a complete list of all these writings and of the genre marks, see [3]. All the proper nouns were so marked, as they may not correctly contribute to any statistical measurement of the written language except for these proper nouns themselves. These nouns were marked in the original texts by research assistants with enough command of the language to make correct judgment. Anything else, including punctuation marks of all sorts, in the texts were properly processed. Every sentence, including some vocative phrases, was numbered within the writing piece quite mechanicaly, though occasionally it was necessary for specialists to make certain judgment for segmenting sentences.

The Code System

The Chinese standard telegraphic code system includes some 9500 codes for Chinese characters. A code consists of a set of 4 digits, which represents one Chinese character. Among those 9500, 5231 have been used.

Statistics

Statistical analysis of this million-character file can be found in [4]. Some additional statistics are provided here. Fig. 1 shows the 10 most frequently used characters with their frequencies. These 10 characters occupy 17.1% of the total amount. Fig. 2 is a table of character frequencies vs. the number of character types. Fig. 3 shows the cumulative percentage of character occurrences as a function of the number of character types (in descending order of frequency). It indicates, for example, only 92 characters represent 47% of the entire data. There are 1170 characters each of which are used more than 100 times and they occupy 92.8 % of the whole data.

Character	Frequency
的	46531
一	18077
是	17874
了	16390
不	16138
我	12827
在	11096
人	11057
有	10717
他	10332

Fig. 1. List of High Frequency Characters

Frequency	No. of Character Types
- 10001	10
10000 - 5001	13
5000 - 3001	32
3000 - 2001	37
2000 - 1001	106
1000 - 501	176
500 - 301	208
300 - 201	191
200 - 101	397
100 - 81	150
80 - 61	230
60 - 41	294
40 - 21	574
20 - 11	563
10 - 1	2250

Fig. 2. Frequency Distribution of Chinese Character Types

CHINESE CHARACTERS VS. KANJI

Chinese characters were imported into Japan sometime in the 5th century. Since then, they have been extensively used with a few additional characters created in Japan (this modified set of Chinese characters is called "kanji"), although hiragana and katakana (two sets of pure Japanese characters with their origin also in the forms of Chinese characters) were invented early in the 9th century.

"Chinese characters for daily use" established by the Ministry of Education for modern Japanese includes a 1850 kanji set, however several thousand more are still in use especially for proper nouns. The Japanese Industrial Standard (JIS) "Code of the Japanese Graphic Character Set for Information Exchange (C6226)" established in 1978 includes a 6349 kanji set, hiragana, katakana, Roman alphabet, Greek letters, Russian letters and other symbols. The kanji set is grouped into 2 levels, the first level a 2965 kanji set and the second level a 3384 kanji set. This means some 3000 kanji are considered to be enough for basic information exchange in Japanese. In this experiment, the kanji printer system T4100 (Syowa Zyoho, Co., Ltd.) was used. A total of 8182 characters was available for this printer including 7360 kanji, hiragana, katakana, Roman alphabet, and other miscellaneous symbols. The system was developed 5 years before the establishment of JIS C6226.

As mentioned before, the million-character file included 5231 different Chinese characters. Among them, 295 were found to be unprintable (because they were not found in the T4100 system). The fonts of those 295 characters were designed and incorporated into the T4100 system. Later, when JIS C6226 was established, some of those 295 characters were found in the second level of the kanji set, namely 龔(frequency 773), 鬥(581), 黨(563), 隴(345), 鐙(343), 鈇(189), 餘(178), and 絲(158). Fig. 4 shows the frequency of the remaining 287 characters. Their total frequency numbers 1100, which is 0.1% of the million-character file. This fact indicates that Chinese characters and kanji still overlap closely in modern Chinese and Japanese. (It should be noticed that the simplified Chinese characters are out of this scope since they did not exist at the so-called May 5 Movement period.)

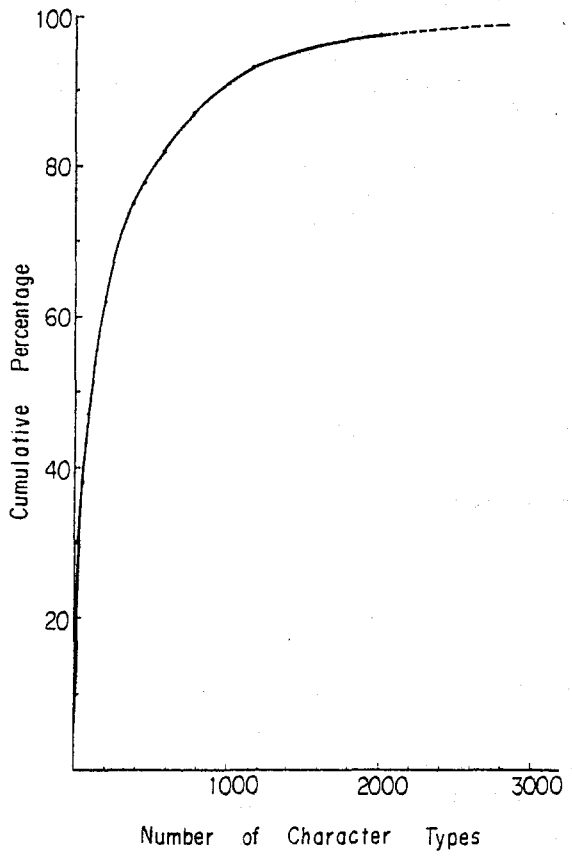

Fig. 3. Cumulative Percentage of Character Occurrences as a Function of the Number of Character Types

THE CONCORDANCES

Besides the text itself, the Princeton million-character file contained information on the title, the author, the sentence numbers, and other miscellaneous editorial symbols (such as

Frequency	No. of Character Types
554	1 (狗)
228	1 (桌)
134	1 (碰)
128	1 (鯛)
100 - 51	7
50 - 31	7
30 - 21	8
20 - 11	21
10 - 5	37
4	11
3	34
2	41
1	117

Fig. 4. Frequency Distribution of Chinese Characters which are not Found in the Kanji Set

marks to indicate proper nouns). Extensive work had to be done to interpret and reform editorial symbols. Fig. 5 shows the edited text sentences from the million-character file. After this editorial step and incorporation of Chinese character fonts to the T4100 kanji printing ststem, the concordance compilation process was started. Since we have had experience with the automatic compilation of one-million line concordances in Japanese [5], not many technical difficulties were encountered, except some malfunctions of our old kanji printer. Discussions on the salient features of those Chinese concordances follow.

Key Words

KWIC index style has been adopted as the form of Chinese concordances, since it is one of the most fundamental styles for computer-generated concordances. Because there is no clear segmentation of words in Chinese, and because one character represents a fairly sizable amount of information, each character was chosen as a "key word". Furthermore, no elimination of "non-key words" were made. Every character (including punctuation) was chosen as a key character. In this sense, the concordance may be named as "All characters in context" index. Consequently, one million character data required one million lines of index.

Contexts

One of the deficiencies of the KWIC index style is that the context each line can show is limited to its line length. We could afford 55 characters for the context. Since one or two Chinese characters represent a word, this length can accommodate more than 30 words of information in English.

Reverse Sorted Index

Two types of KWIC index have been produced. One is for the normal type, in which all lines are sorted in the ascending order of the Chinese standard telegraphic code of key characters (plus 7 succeeding characters). Fig. 6 shows an example page from this type of index. The other is the so called "reverse sorted" index. The major key for this type is the same as that of the normal type. The minor sort keys are, the characters immediately preceding the major key. Thus all lines for one key character are listed in the ascending order of the code for the character immediately preceding the key character and so on. Fig. 7 shows an example page from the reverse sorted concordance.

CONCLUDING REMARKS

The two sets of modern Chinese concordances can be reached at the National Inter-university Research Institute of Asian and African Languages and Cultures, Tokyo University of Foreign Studies. It should be noted that a concordance of one million lines amounts to over 25,000 pages (actually it counts for 27,341) or 50 volumes of a 5cm-width paper file. Before printing the whole index, engineers recommended linguists to use COM technique, but in vain. A microfiche version should have been produced for portability. Analysis of the concordances have just got off the ground. The resulting papers are expected to follow.

作家	作品	文番号	原　文
艾蕪	石青	1	早上太陽仍像往天一樣，把晴美的陽光抹上滿峡的樹林，叫帶露的樹葉草葉都亮得耀人的眼睛。
		2	只是石青嫂子的心上却陰暗極了，陰暗得像夏季烏雲滿佈的天空一樣，随時都会雨点似的落下泪来。
		3	看見屋裡踢倒的板凳，打爛的燈，再看見門前地裡一片乱踏的足跡。
		4	菠菜的葉子，踩来変成爛泥；番茄踩成一灘一灘的紅漿。
		5	那些紅漿很使石青嫂子疑心，怕是夜来扭扯的時候，他身上流出来的血。
		6	对河山腰上的汽車公路，一乘長途汽車馳過以後，便比平日還要静寂，簡直静寂得可怕，滿山禿露的乱石，在陽光下面更加顕得蒼老醜陋，彷彿一些生癩疤的禿頭似的。
		7	人工鑿過的公路，隠蔵在乱石裡面，一種原始的荒涼的気氛，越発強烈地流露出来。
		8	有石青的時候，她従来没有感到過她這間山峡中唯一的茅屋是孤独的，寂寞的，可怕的。
		9	她只覚得面臨小河，背靠山嶺的一片斜坡，給予她無限的繁忙和労碌。
		10	她終天頭上包著一張藍布帕子，不是拿鋤頭挖地，鐮刀割草，就是手腕上掛個籃子，採摘什麼東西。
		11	晚上星子都現在山峡的高空了，樹林茅屋全隠蔵在軽霧裡面，小的孩子，坐在門前哭著喊媽的時候，她還在地裡摘著苦瓜豇豆或是茄子辣椒，準備明天一早挑到五里以外的鎮上去売，好換点米回来。

Fig. 5. An Example from the Edited Text

，我国生産資料所有制的社会主義革命，已経基本完成。 這 是一個巨大的変革。但是，僅有経済制度上的這個変革； 劉少奇在北 5 6
国是一個社会主義的大国，但又是一個経済落後的窮国， 這 是一個很大的矛盾。要使我国富強起来，需要幾十年艱苦 毛沢東関于 6 0 2
的詩就是一種不負責的東西了，不負責的東西是好的嗎？ 這 是一個很重要的問題，所以，第一種主張，就側重在這種 聞一多詩与 1 8
和人民都在実際上懂得這個問題。為什麼現在又有人覚得 這 是一個新問題呢？這是因為過去国内外的敵我門争很尖鋭 毛沢東関于 1 3 6
如何把地下発掘的資料，与伝下来的紀録資料連綴起来。 這 是一個普遍的考古問題。也是在中国区域，考古家所面臨 李済 再談 1 9 8
倒是蹲在家裡坐冷板凳好些。那第三科科長没有管理他， 這 是一個沉悶而少話説的青年人，油黒的面孔上生着幾顆面 沙丁 代理 2 7
頭上坐着，沐浴着藍色的霧，漸漸地感到了老年的沉重。 這 是一個没有月色的初夜。没有遊人。衰草裡也没有蟋蟀的 何其芳遲暮 7
受人注意而其後竟受到揺頭，這便是一個解釈，不革命。 這 是一個真実，我們応該有勇気来承認這真実，承認這失敗 茅盾 従牯 1 1 3
們要進行大規模的建設，但是我国還是一個很窮的国家， 這 是一個矛盾。全面地持久地属行節約，就是解決這個矛盾 毛沢東関于 5 8 4
中一般。三面都是山，像半個環児擁着；人如在井底了。 這 是一個秋季的薄陰的天気。微微的雲在我們頂上流着；岩 朱自清温州 4 0
底変局。中国人本来城裡人，到此時忽然成為郷下人了。 這 是一個空前底変局。這是中国人所遇到底，一個空前底挫 馮友蘭辨城 6 8
関於正確処理人民内部矛盾的問題， 這 是一個総題目。為了敘述的方便，分為十二個小題目。在 毛沢東関于 1
表表示出来的。我們仍以下面這個三段論定理為例ＱＱ。 這 是一個複合命題，它有三個原始命題Ｑ。這三個原始命題 金岳霖論推 5 5 2
，這自従工業革命之後在西洋所発生的那一套生活方式， 這 是一個豊裕的経済。我並不覚得自己配談西洋文化，我缺 費孝通中国 1 7 1
没看清那人是誰，已経把那人抓過来摔在地上。他断定了 這 是一個賊。「多多頭。打死我也不怨你，只求你不要説出 茅盾 春蚕 3 2 5
来，坐在炕沿：「柴火焼没了。」女人説，瞅老趙一眼。 這 是一個跟他吃尽千辛萬苦，也不抱怨的好心眼的小個子女 周立波暴風 4 8 0
們開始看到了現実主義的創作要求底多方面的発展現象。 這 是一個重要的転換期，因為，在社会学的意義上説，思想 胡風 文藝 2 7
放在枕辺的皮包拿過来，従中取出一個摺子，他叫她瞧。 這 是一個銀行的摺子，浄存一欄上，七千五百二十三元。外 張天翼報復 3 4 5
従前採用的黨内鬥争方法叫做「残酷鬥争，無情打撃」。 這 是一個錯誤的方法。我們在批評「左」傾教條主義的時候 毛沢東関于 9 9
瞥上那朵「蚕花」，跟鵝毛一塊摔在「蚕簞」的辺児上。 這 是一個隆重的儀式。千百年相伝的儀式。那好比是聲師典 茅盾 春蚕 2 4 0
不巧撞着了一個巡路的小長毛，当時没法，只好殺了他， 這 是一個「結」。然而従老通宝懂事以来，他們家替這小長 茅盾 春蚕 3 3
応」起来的。韓愈談他自己做古文，「惟陳言之務去。」 這 是一句最緊要的教訓。語言跟着思想情感走，你不肯用俗 朱光潜咬文 8 6
的片子就是今天晚上…您不是早就説過？…」阿宝摸不清 這 是一回甚麼事，粉臉太太驟然添上了一臉怒色，円胖的鼻 王統照小紅 2 0 9
旁辺還有一株牽牛藤在晩風里微微舞動它的柔軟的腰肢。 這 是一幅静的，美麗的，幻想的図画。我不覚痴痴地望着它 巴金 窗下 1 5 0
這 是一張尺多寛的小小的横幅，馬孟容君画的。上方的左角 朱自清温州 1
初見你的時候，那就老得多了。」這是一件很巧的事。「 這 是一所三楼三底両夾廂的上海式楼房。凰二爺住楼上的客 梅蘭芳舞台 4 3
主義，似乎行政事務做得多了，所以就放鬆了思想工作。 這 是一方面的理由，但不是全面的理由，甚至不是主要的理 周揚 整頓 1 7 7
麼理智都要躲避的，他未経考慮一下就和你訂婚了。可是 這 是一時的，你記着，這是一時的，過後，熱情一冷了下去 張天翼報復 2 3 9
未経考慮一下就和你訂婚了。可是這是一時的，你記着， 這 是一時的，過後，熱情一冷了下去，他会覚得他自己滑稽 張天翼報復 2 3 9
抗侵略和反抗圧迫的人民戦争。人民戦争必将蓬勃発展， 這 是一條客観規律，既不以美帝国主義者的意志為転移，也 林彪 人民 2 3 6
大衫，帯上門出去。沿着荷塘，是一條曲折的小煤屑路。 這 是一條幽僻的路；白天也少人走，夜晩更加寂寞。荷塘四 朱自清荷塘 5
地築堰的臂，要将他的妻孥奴子推到不可知的安全地方。 這 是一條悠長而単様的路。宿羊山雖遥遥在望，但它的倩影 蕭乾 宿羊 2 9
当什麼都好了，就可以不費気力享受現成的幸福生活了， 這 是一種不実際的想法。我国少数民族有三千多萬人，雖然 毛沢東関于 3 7 7
他並不覚得這有什麼説不過去；有時候擋上他，他還覚得 這 是一種優越，那些拉破車的根本就用不上竜石燈。現在， 老舎 駱駝 4 5 6
主要的創作力量移用到文藝行政工作方面去的危険傾向， 這 是一種喧賓奪主，殺雞求卵的辦法。没有作家，没有作品 周揚 整頓 1 8 2
翼戴政府，擁護政府，天下清平，門第亦同享安泰之楽， 這 是一種大気度。他們在政治上常抱一種領先的姿態。西漢 銭穆 中国 3 0 0
以，第一種主張，就側重在這種宣伝的効果方面，我想， 這 是一種対於詩的価値論者。好些人念一篇詩時是不理会他 聞一多詩与 1 9
有了空前広泛的発展，許多民族独立国家已経建立起来。 這 是一種巨大的世界力量。一切被圧迫民族的独立自主的要 劉少奇在北 3 3
従朝鮮帰来的人，会知道你正生活在幸福中。請你意識到 這 是一種幸福吧口巴，因為只有你意識到這一点，你才能更 魏巍 誰是 1 1 6
的作家；假如你為小資産階級訴苦，便幾乎罪同反革命。 這 是一種很不合理的事。現在的小資産階級没有痛苦麼？他 茅盾 従牯 1 3 4
領導知識分子進行文化建設，我們就天然地不会犯錯誤。 這 是一種很危険的想法。而在有些地方，我們有一些同志正 周恩来関于 3 5 2
様的死一次。這已超過了忍受的問題，在他們，十分覚得 這 是一種恥辱。這裡一直不会旱下去的，但是，他們等不得 端木良大江 8 9
6 6 3 8 這 （這是一這是）

Fig. 6. A Page Example from the Chinese Concordance (Normal Style)

動的唯心論哲学家，他們也有一套唯物論与唯心論的統一 的 理論。我們前面已経批評到的黒格爾，就是突出地站在唯 賀麟 論唯 1 5 3
之愛」，自従人類分化成為階級以後，就没有過這種統一 的 愛。過去的一切統治階級喜歓提倡這個東西，許多所謂聖 毛沢東在延 3 6 4
成唯心論統一於唯物論，而不是相反。由於這個対立統一 的 原理意義特別深遠而豊富，完全可以適用於説明唯心論与 賀麟 論唯 1 4 7
事物的矛盾法則，即対立統一 的 法則，是唯物辯證法的最根本的法則。列寧説：「就本来 毛沢東矛盾 1
，我們可以総起来説幾句。事物矛盾的法則，即対立統一 的 法則，是自然和社会的根本法則，因而也是思維的根本法 毛沢東矛盾 4 8 7
問題，得到了如下幾点認識。第一，要認真貫徹対立統一 的 唯物辯證法原理，必須承認唯心論与唯物論是可以相互融 賀麟 論唯 1 4 5
鬥争的統一，於重大現実毛主席所提出的関於鬥争与統一 的 基本原理，却有偉大的啓示性，完全可以応用来討論唯物 賀麟 論唯 1 2 5
由各個有関部門直接負責，而由中央宣伝部負責進行統一 的 監督。中央宣伝部応該経常検査各部門各地区執行中央関 周恩来関于 3 6 5
様黒也是空的。現在是「陽春白雪」和「下里巴人」統一 的 問題，是提高和普及統一的問題。不統一，任何専門家的 毛沢東在延 2 7 9
死。所以便懶著不動了。但是終於還是得走，想得到萬一 的 機会，假使碰見水呢？於是従炎熱裡向前走。重傷的随走 端木良大江 5 3
是水底月，鏡中天，但在自己的心裡不能否認総含有萬一 的 希望。因此，我自従十四歳以後便不願従速訂婚。我的父 郭沫若黒貓 1 1
仰一仰，像画家設完一層色那麼退後看看。然後，又逐一 的 転開，把另一面的魚們擺斉。又往後仰身端詳了一番，回 老舎 黒白 2 0
三個女児的悲慘結局，你的懐疑慢慢会変成惆悵。在園丁 的 樸実言語中，伝説中的古怪老頭和他的女児重新復活過来 師陀 果園 1 1 0
臉細眉，白浄皮色，太太模様的女人，年紀不過二十六七 的 光景。我不会抽烟，我也没有買什麼東西，我只来她換一 巴金 生与 8
声音乱叫了。但是那位作調人的警察却冷笑，扳著陳老七 的 肩膀道：「我勧你少找点麻煩罷。到那該，中什麼用。你 茅盾 林家 8 2 4
朱三太的上門討利息，他記起還有両注存款，橋頭陳老七 的 二百元和張寡婦的一百五十元，総共十来塊銭的利息，都 茅盾 林家 2 0 2
略顕渾濁，有出山泉水的意思。若溯流而上，則三丈五丈 的 深潭皆清澈見底。深潭中為白日所映照，河底小小寛石子 沈従文辺城 5 6
南京大学教授和我談話的地方，即離開左面的断崖数十丈 的 地方，我的確看到有一根不很大的石條伸出在空中，照相 豊子愷廬山 2 6
可以跨過似的。然而実景中並没有石條，只是相距若干丈 的 両個断崖，我們所登的便是左面的断崖。我想：這地方叫 豊子愷廬山 2 1
指著他們誇耀似地対我們説：「這両位高僧，是我們方丈 的 師兄，年紀都快八十歳了，是従城裡某公館裡回来的。」 郁達夫遲桂 4 0 6
抱著羊竿，哀喚著「天妃」的慈号。我的心舟在起落萬丈 的 思潮中震蕩時，母親。縦便你在萬里外，写到「母親」両 謝婉瑩寄給 1 6
域基礎，同時却也有它的社会基礎。所以，你如把「張三 的 爸爸巴」説成「張三他爸爸巴」，或者你不説「他還没来 董同龢国語 7
生皺著高鼻上的皮，揺頭継続説：「撒爛汚。十個阿孀三 的 血，有九個不清潔。毒来些，齷齪来些。要不得，要不得 呉組湘官官 1 6 4
郷下你那些田，早早脱手得好。自従改了民国，接二連三 的 打仗，倒嘗有一年間過，把地面上糟塌得不成様子，中間 張愛玲金鎖 5 7 4
的声音向覚民問道。「我剛剛看見抬傷兵進城，接二連三 的 ，不暁得有多少，」覚民激動地用顫抖的声音説，「真可 巴金 家 2 1
玩児去，誰忍耐站在先生書桌前，晃著身子，背早上上 的 多難的書？啊，飛。不是那在樹枝上矮矮的跳著的麻雀児 徐志摩想飛 7
綫上領導前進者，総是少数特殊分子，遺落在後追随不上 的 ，依然混同一色，那纔是社会整体之真骨幹。結果詩人， 銭穆 中国 2 2 7
…。假使没有了頭顱，却還能做服役和戦争的機械，世上 的 情形就何等地醒目呵口可，這時再不必用什麼制帽黙章来 魯迅 春末 5 5
是共産的事実，不是言論。欧洲之所以駕乎我們中国之上 的 ，不是政治哲学，完全是物質文明。因為他們近来的物質 孫逸仙三民 4 2 5
方面的研究，成為革命政黨正確地決定其政治上和軍事上 的 戦略戦術方針的重要方法之一，是一切共産黨人都応当注 毛沢東矛盾 3 5 1
。社会主義已経成為世界体系。占人類総数三分之一以上 的 社会主義国家，已経組成了以蘇聯為首的社会主義陣営。 劉少奇在北 3 0
得到的，有関這両個基本問題的答案及解釈，證実了以上 的 説法。整理田野考古発掘所得的資料，一個最迫切的問題 李済 再談 1 9 7
這時還不改変策略，它将坐困于為数至少在二十萬人以上 的 義和団民的包囲之中，而其結果就是這個王朝的統治被粉 翦伯賛義和 1 2 2
，聯保主任跑来報告，説是索橋辺已経扣留下二十個以上 的 災民了。才一走進屋裡，還来不及把醃肉掛向籬壁的竹釘 沙丁 代理 2 0 0
弾会怎様地鑽進他底身体，他也感到一種超乎圧制力以上 的 恐怖。電燈光閃閃地刺痛他底眼睛。「這燈光。」他煩躁 巴金 家 7 8
是人民大衆呢？最広大的人民，占全人口百分之九十以上 的 人民，是工人，農民，兵士和城市小資産階級。所以我們 毛沢東在延 1 3 8
的革命的功利主義者，我們是以占全人口百分之九十以上 的 最広大羣衆的目前利益和将来利益的統一為出発点的，所 毛沢東在延 2 7 5
的恋愛已経失去了荘厳的時候，随時随地可以有接吻以上 的 行為的。去年是不是我和你当了你未婚夫的面互相擁抱， 張天翼報復 2 3 5
和他們在江北的地都管計在内，他們三家都有一千塊以上 的 好地，係通和黄土包子還不管在内。街裡的福来徳焼鍋， 周立波暴風 1 4 6
外。単純的過程只有一対矛盾，複雑的過程則有一対以上 的 矛盾。各対矛盾之間，又互相成為矛盾。這様地組成客観 毛沢東矛盾 3 5 9
在小屯的建築遺址中，地基部份的夯土，有夯至十層以上 的 ；好些地方，把夯過了的土毀了再夯大力。這一営造方式 李済 殷墟 5 7
那人頭後面突出一個硬硬的小円餅，顕然是一個中年以上 的 女人，挽著髪髻彫吉。她的心往下一沉。她知道那是金花 張愛玲秧歌 3 3 2

4 1 0 4 的 （一的一上的）

Fig. 7. A Page Example from the Chinese Concordance (Reverse Sorted Style)

REFERENCES

1. Kierman, F.A. and Barber, E.: "Computers and Chinese linguistics", Unicorn, No. 3 (1968)
2. Boltz, W.G., Barber, E. and Kierman, F.A: "Progress report on Pai-hua-wen computer count and analysis", Unicorn, No. 7, pp. 94-138 (1971)
3. Hashimoto, M.J., et al.: A grammatical analysis of the Princeton million-character computer file", Bulletin of the Chinese Language Society of Japan, No.222, pp. 1-16,36 (1975)
4. Hashimoto, M.J.,: "Computer count of modern Chinese morphemes", Computational Analysis of Asian and African Languages, No. 7, pp. 29-41 (1977)
5. Uemura, S.: "Automatic Compilation and Retrieval of Modern Japanese Concordances", Journal of Information Processing, Vol. 1, No. 4, pp. 172-179 (1979)