现代汉语常用词词频词典

（音序部分）

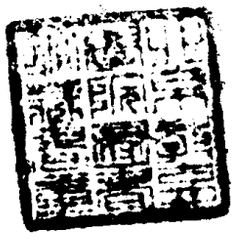
刘　源　梁南元　王德进　张社英
杨铁鹰　揭春雨　孙　伟　编

学林出版社

现代汉语
常用词词频词典

（音序部分）

编者：刘　源　梁南元　王德进　张社英
杨铁鹰　揭春雨　孙　伟



学报出版社

内 容 简 介

本书是从国家科研项目“现代汉语词频统计”的部分统计结果中选录的。在国家科委、国家技术监督局领导下，该项目聘请了钱伟长、王力等16位专家作技术顾问。

全书收录了《现代汉语词典》、《辞海》、《汉英词典》等23部影响较大的工具书中的常用词条，充分反映了我国使用汉语词条的基本情况。本书为音序表。

本书是国内外规模最大、被统计材料分科最多、时间分布最长、且首次用计算机自动进行的一次汉语通用词频统计。它具有背景干扰小、通用性强、分词标准一致、统计精度较高等显著优点，是中文信息处理的工具书。

本书对信息产业的建立与交流，对中外语言文字的传播与转换，对汉语语言理解、汉语教学、语言学研究、情报检索、速记、机器翻译、人工智能、盲文、手语编制、办公室自动化等中文信息处理技术的开发，都有重要的意义。它也是计量语言学的基础。目前已应用于汉字编码、国家标准词库的建立等领域。

现代汉语常用词词频词典（音序部分）

编者： 刘 源 梁南元 王德进 张社英
杨铁鹰 揭春雨 孙 伟

责任编辑： 廖寿琪

*

宇航出版社出版
地址：北京和平里滨河路1号
邮政编码：100013
宇航出版社激光照排室排版
各地新华书店经销
天津静·胶印厂印刷

*

开本：787×1092 1/16 印张：83.5 字数：2084千字
1990年6月第1版第1次印刷 印数：1—3000册
ISBN 7-80034-302-2/H·031 定价：51.5元

前　言

《现代汉语常用词词频词典》终于出版了。这部巨著是根据“现代汉语词频统计”科研项目统计结果编纂的。“现代汉语词频统计”开始于1981年，于1986年6月30日通过国家级鉴定，历时六年，北京航空航天大学等10个单位前后有数百人参加了这项规模巨大的工程。《现代汉语常用词词频词典》只收录了这次词频统计的总结果，各分时期、分科结果尚不包括在内，但已洋洋数百万言。它的出版，实可告慰为之呕心沥血、奋斗经年的科研工作者。

汉语是世界上使用人数最多的语言，也是历史最悠久的语言之一。汉语不仅对汉民族的感情交流，而且对中华民族的统一都作出了巨大贡献。但是由于汉语的方块汉字的特点，以前对它进行定量研究甚少，对词汇的定量研究则更少一些。这次词频统计是世界上迄今为止规模最大的汉语词频统计，被统计出词条达七万之多，对二千余万汉字（包括标点符号）进行了统计，被统计材料遍及社会科学和自然科学的政治经济、历史哲学、新闻报道、文体生活、文学艺术、建筑运输、农牧副渔、电子轻化工、重工业、基础科学等十类学科。这项基础性的工作将对汉语语言研究及汉语信息处理产生深远影响。它也说明了中国的汉语信息处理在许多方面已达到或超过国际先进水平。

作为“现代汉语词频统计”的十六名顾问之一，以及作为这个科研项目的鉴定委员会主任委员，我由衷高兴的看到《现代汉语常用词词频词典》的出版，并且希望有更多的汉语信息处理新成果问世。

钱伟长

1987年10月

Forword

"Dictionary of Usage Frequency of Modern Chinese Words," by all standards is a monumental work. It was compiled on the basis of "A Statistical Survey on the Usage Frequency of Modern Chinese Words," a research project which started in 1981 and passed a state-level evaluation on 30 June, 1986. The project lasting six years was jointly undertaken by 10 institutions including the Beijing University of Aeronauties and Astronautics, involving hundreds of people. Ineluded in the dictonary is the general result of the survey but not data for specific historical and subjects. Never the less, it runs to millions of words. Publication of the dictionary is the greatest consolation to those who have devoted so much and for so long for the time.

Chinese is a langauge spoken by the greatest number of people in the world; it is also one of the language with the longest history. It has contributed immensely not only to communication among the Han people but also to the unification of the Chinese nation. However, due to the non-phonetic character of the chinese language, little effort was made in the past on a quantitative study of the language, let alone its words.

The statistical survey was the biggest in scale in the world on the usage frequency of Chinese words. It covers more than 70000 words and over 20 million characters (including punctuation marks). Data have been gathered from many fields in social and natural sciences, including politics, economics, history, philosophy, news reporting, art and culture, sport, architecture, transport, agriculture, electronics, light and heavy industries, chemical industry and basic sciences. The groundwork laid by the survey will have a far-reaching influence on the study of the Chinese language and information processing using the language. It indicates that Chinese-language information processing in China has reached or even surpassed international advanced levels in many aspects.

As one of the 16 consultants for the "A Statistical Survey on the Usage Frequency of Modern Chinese Words" project and as the president of the Evaluation Committee for the project, I feel so happy to see the publication of "Dictionary of Usage Frequency of Modern Chinese Words" and hope that the dictionary heralds the coming of more research result on chinese-language processing.

Qian Wei-chang

Oct. 1987

まえかき

《現代漢語常用ことばのフリケンシー辞典》はもう出版された。この大著は“現代漢語言葉フソケンシー統計”の科学研究項目の統計結果に基づいて編さんされたのである。“現代漢語言葉フリケンシーの統計”の工作は1981年から始まり,1986年6月30日に国家級の評定を通過した,六年を経て,北京航空航天大学などの10部門は前后,数百人がこの巨大な工程に参加しな。《現代漢語常用言葉フリケンシー辞典》はこんどの言葉フリケンシーの統計の総結果なヒニナ採用して各分時期と分科の結果をふレまれなソがもうようよう数百万言葉になつた,これガ出版されたのはこの仕事のために胸襟を開ソて刻苦奮斗する科学関系者になぐさめる。

漢語は世界で使う人数がいちばん多い語言である,暦史ガウをていちばん悠久なる語言のひとつである。漢語は漢民族の感情の交流だけでなく,中華民族の統一に対して巨大な貢献をするものであゐ。しかし漢語は方形の文字の特点によてて,むかひはこれに定量的に研究することは少なかつたし,語彙に定量的に研究すゐこともつと少なかプたこんどの言葉フリケンシーの統計はソままで世界で最大の規模の言葉フリケンシーの統計である。統計された言葉は七万に達した,句読点を合んで二千万あまり漢字に対して統計した、統計された材料は社会科学と自然科学の政治経済、暦史哲学、新聞報導、娯楽体育生活、文学芸術、建築運輸、農業、副業、牧畜業、漁業、電子工業、軽工業、化学工業、重工業、基礎科学など十種類の学科にゆきゐたる。この基礎的な工作は漢語の研究ゐよび漢語の情報処理にふかく影響を与えている。このことは中国の漢語の情報処理が各方面でもう国際先進の水平に達してあるいは超過したことを説明した。

“現代漢語言葉フリケンシーの統計”の十六名顧問のひとつとして,あるいはこの科学研究項目の評定委員会の主任委員として,《現代漢語常用言葉フリケンシー辞典》の出版することを見て心から喜び,そしていそ;多く漢語処理の新しい成果を世に送り出すことを望む。

Préface

Le "Dictionnaire de la fréquence de l'emploi des mots usuels du chinois contemporain" a vu le jour. L'ouvrage a été rédigé d'aprés un projet de recherche scientifique déjà réalisé, intitulé "Statistiques de la fréquence de l'emploi des mots du chinois contemporain". Le projet, qui rassemblait plusieurs centaines d'experts, venus de l'Université aéronautique et astronautique de Beijing et de dix autres unités, a commencé en 1981. Au bout de six ans de recherches, il a été officiellement approuvé le 30 juin 1986. Bien que le dictionnaire se borne à inclure les mots généraux récensés, en excluant les statistiques du vocabulaire spécialisé faites à chaque stade et pour chaque discipline, son vocabulaire atteint plusieurs millions de caractéres chinois. Le parution de ce dictionnaire a certainement rempli de joie les experts qui lui ont consacré plusieurs années d'études.

Le chinois, une des plus anciennes langues du monde, est parlé par le plus grand nombre de personnes au monde. Il est non seulement parlé par les Hans, mais il a joué aussi un grand role dans l'unification de la nation chinoise. La structure complexe du caractére chinois qui s'inscrit dans un carré a rarement fait l'objet d'une analyse quantitative, et le vocabulaire non plus. Le recensement de ces mots, le plus important de chine par son ampleur, porte sur 70000 articles et plus de 2 millions de caractéres (y compris les signes de ponctuation). L'ouvrage comprend dix disciplines scientifiques: science politique et économie, histoire et philosophie, jounalisme, sports, culture, art et littérature, architecture et transport, agriculture, élevage, pêhce et occupations auxiliaires, électronique, chimie et industrie légère, industrie lourde, et sciences fondamentales. Il est certain que ce travail de base aura une influence profonde sur la recherche sur la langue chinoise et sur son traitement informatique. En outre, la réussité du projet montre que la Chine a atteint, et même dépassé les normes internationales dans plusieurs domaines, en ce qui concerne le traitement informatique du chinois.

En tant qu'un des 16 conseillers du projet de recensement et le président de la commission de vérification du projet, je suis heureux de constater la parution du présent dictionnaire, en espérant voir paraître d'autres ouvrages en matiére de traitement informatique de la langue chinoise.

Qian Weichang

octobre 1987

Vorwort

"Das moderne chinesische Häufigkeit-Wörterbuch" ist nun herausgegeben. Dieses große Werk ist auf der Grundlage der Ergebnisse des Forschungsprojektes "Statistik über die moderne chinesische Häufigkeit" zusammengestellt worden. Das Projekt "Statistik über die moderne chinesische Häufigkeit" wurde im Jahre 1981 in Angriff genommen und nach sechs Jahren am 30. Juni 1986 vom Staat begutachtet. Einige hundert Personen aus elf Einheiten wie der Beijinger Universität für Luft und Raumfahrt arbeiteten an diesem großen Werk. "Das moderne chinesische Häufigkeit-Wörterbuch" enthält nur das Endergebnis der Statistik über die chinesische Häufigkeit, die Ergebnisse der Statistik über die chinesische Häufigkeit in einzelnen Perioden und einzelnen Disziplinen nicht einbegriffen. Das Wörterbuch hat einige Millionen Schriftzeichen. Mit der Herausgabe des Wörterbuchs fühlen sich die Wissenschafter getröstet, die jahrelang dazu große Anstrengungen unternommen haben.

Die chinesische Sprache ist eine Sprache, die von dem größten Teil der Weltbevölkerung gesprochen wird, und zählt zu den altesten Sprachen. Sie trägt nicht nur zum Gedankenaustausch der Han-Nationalität sondern auch zur Einheit der chinesischen Nation bei. Infolge der Besonderheiten der chinesischen Schriftzeichen wurden wenige quantitative Forschungen der chinesischen Schrift und noch wenigere quantitative Forschungen der chinesischen Wörter vorgenommen. Diesmalige Statistik über die chinesische Häufigkeit ist die größte Statistik in diesem Bereich. Es wurden insgesamt über 70000 Stichwörter und über 20 Millionen einzelner Wörter (einschließlich der Interpunktionen) zusammengefaßt. Die Statistik erfaßte Ausdrücke in über zehn Disziplinen der Gesellschafts-und Naturwissenschaften wie Politik, Wirtschaft, Geschichte, Philosophie, Journalistik, kulturelle und sportliche Betätigungen, Literatur und Kunst, Bauwesen, Transportwesen, Landwirtschaft, Viehzucht, Nebengewerbe, Fischerei, Elektronik, Leichtindustrie und chemische Industrie, Schwerindustrie und Basiswissenschaften. Diese grundlegende Arbeit wird einen tiefen Einfluß auf die Forschung der chinesischen Schrift und Datenbearbeitung der chinesischen Schrift ausüben. Dies beweist, daß China in vielen Sektoren der Datenbearbeitung der chinesischen Schrift das Weltniveau erreicht bzw. übertroffen hat.

Als einer der sechzehn Berater der Redaktion des "modernen chinesischen Häufigkeit-Wörterbuchs" und Vorsitzender der Gutachtergruppe dieses Forschungsthemas freue ich mich auf die H ausgabe dieses Wörterbuchs und hoffe, daß immer mehr neue Ergebnisse in der Datenbearbeitung der chinesischen Schrift erzielt würden.

Qian Weichang

Oktober 1987

ПРЕДИСЛОВИЕ

Вышел в свет "Частотный словарь наиболее употребительных слов современного китайского языка." Словарь составлен на основе итогов "Статистического исследования частоты слов современного китайского языка", начатого в 1981 году и прошедшего государственную экспертизу 30 июня 1986 года. На протяжении шести лет в этом огромном труде приняли участие сотни научных работников Пекинский Азро- космический университет и других 10 организаций. В " Частотнчй Словарь" объёмом в миллионы иероглифов вошли только общие итоги данного статистического исследования, не включая данные исследований по периодам и дисциплинам. Его издание служит наградой научным работникам, отдавшим все силы многолетним исследованиям.

Китайский язык—— один из древнейших языков мира, занимает первое место в мире по числу говорящих на нем людей. Китайский язык не толвко внес огромный вклад в общение ханьцев, но и в дело об "единения китайской нации. Однако вследствие особенностей иероглифического письма количественное исследование китайского языка, а тем более его лексики проводилось крайне недостаточно. Данное лингво-статистическое исследование китайского языка является крупнейшим по масштабам в мире. Более 70 тысяч словарных статей, вошедших в словарь, получены в результате обработки текстов об" емом более чем в 20 мнллионов китайских иероглифов (включая знаки препинания) в областях общественных и естественных наук, как- то: политэкономии, истории, философии, журналистики, культуры, спорта, литературы и искусства, архитектуры, транспорта, сельского и подсобного хозяйства, живодтноводства, рыболовства, электроники, легкой, химической и тяжёлой промышленности, фундаментальных наук. Эта фундаментальная работа окажет глубокое влияние на лингвистическое иееледование китайского языка и на обработку информации на китайском языке. Она показывает также, что обработка информации на китайском языке в нашей стране достигла или превзошла во многих отношениях передовой уровень мира.

Как один из 16 консультантов "Статистического исследования частоты слов современного китайского языка " и председатель экспертной комиссии по этой работе я сердечно рад изданию данного "Частотного словаря наиболее употребительных сдов современного китайского языка". Надеюсь увидеть в будущем новые работы по обработке данных на китайском языке.

Цянь Вэйчан

октябрь 1987 года

编　纂　说　明

一　目　的

汉语中有意义的、可以自由运用的最小单位是词，它是汉语研究中非常重要的语言单位，也是计算机汉语信息处理中汉语语言理解、机器翻译等研究的基本单位。汉语词的频度统计对汉语研究、汉语教学、汉语语言理解、机器翻译、汉字编码，甚至对速记、盲文等领域都有重要意义。词频统计也是计量语言学的基础。

汉语由于没有明显的自然形态界限可以作为分词依据；对"词"没有一致认可的定义；同时也因为计算机科学经过了30年迅猛发展才全面应用于汉语信息处理领域，因而长期以来缺少词频统计的结果，尤其缺少大数据量常用词的词频统计结果。这对汉语信息处理和汉语定量研究甚有影响。

出版《现代汉语常用词词频词典》的目的是把科研项目"现代汉语词频统计"（以下简称"词频统计"）中的部分成果经整理后介绍给社会，以方便各个领域的研究人员使用，特别供汉语信息处理学界使用。"词频统计"的全部统计结果有：总结果；社会科学综合结果；自然科学综合结果；4个时期的综合结果；4个时期10类学科的35个分时期分科结果。仅总结果和6个综合结果就有打印纸1万余页。由于力量所限，这次只出版了总结果，——即《现代汉语常用词词频词典》。本词典为音序部分，即：按汉语拼音排列的音序总表以及1至7字词条的音序表。

二　使　用　说　明

2.1 词条

本词典的词条收自"词频统计"软件系统中分词词典的词条。分词词典收取词条130691条，除收取有公认的词外，还有词素、一些词组以及特定的专有名词。专有名词有：人名，地名，国家组织名，书名，民族名等。

有关收取词条的原则见本说明第三节。在出版时删除了人名、书名、一些结合不紧密、使用不固定的词组。有些词条的字不属于国家标准GB 2312－80，因计算机排版困难的原因这些词条也被舍弃。这样本词典的词条有词、词素、词组。主要由于一些词素在特定的语言环境下可以单独出现，也由于自动分词有时发生错误，因此本词典包括词素。有些专有名词和姓氏对某些专门领域的研究有意义，因而本词典中保留了姓氏、地名、国家组织名等。同形异音词条（不包括同形同音不同调词条）有不同的频度。例如，重(zhong4)的统计频度为5146，重(chong2)的统计频度为1094。

本词典收取了"词频统计"总结果中统计频度大于或等于5的词条，低于5次的没有收取。有关词条的分布情况见表1。有关1至7字词条的统计频率见表2。

表 1　词条分布表

词条情况	1 字	2 字	3 字	4 字	5 字	6 字	7 字	总计
分词词典	9199	65891	25352	21699	5124	2446	980	130691
实际统计	7611	46729	11213	9633	1414	675	207	77482
本词典收	5070	31187	5125	4566	382	144	46	46520

表 2　本词典中 1 至 7 字词条统计频率比例表

词条	1 字	2 字	3 字	4 字	5 字	6 字	7 字
使用频率	57.53%	39.25%	1.95%	1.09%	0.11%	0.06%	0.01%
折合字数占总字数比例	39.07%	53.33%	3.97%	2.95%	0.38%	0.25%	0.05%

2.2 词典的编排格式

本词典是按汉语拼音排列的音序总表，以及 1 至 7 字词条的音序表。词表中各栏的意义如下：

1. 序号　词条在词表中的顺序号。
2. 词条　用汉字给出的词条，该词条右上角如有“＊”，则表示它是专有名词。
3. 拼音　词条的汉语拼音。由于目前尚无正式公布的汉语拼音连写方案，故词条中各汉字的汉语拼音分别以拼音 1、拼音 2、……、拼音 7 的形式给出。有些词条有多个声调，在这里一一列出。除轻声字和儿化音的“儿”外，不注变调。为方便计算机处理，在汉语拼音后用 1、2、3、4、5 分别表示声调阴平、阳平、上声、去声和轻声，用 ua 代替 ü，用 ue 代替 üe。用 5 表示轻声是为了方便计算机排序。
4. 频度　词条的统计次数。本词典只收取了“词频统计”中统计次数大于和等于 5 次的词条。
5. 频率　该词条的频度占词表中总频度的百分比。
6. 累频　即累计频率，指已出现词条频度之和占词表中总频度的百分比。

音序表排序时有多个关键字，它们依次是：拼音 1，词条第 1 字，拼音 2，词条第 2 字，……，拼音 7，词条第 7 字。

有关汉字的排序均按照国家标准 GB 2312－80 升序进行。有关汉语拼音中调号的排序按阴平、阳平、上声、去声、轻声的顺序进行。“儿”话音的“儿”的汉语拼音为“r”，并按照“r”排序。

三　“现代汉语词频统计”科研项目简介

本词典是对“词频统计”总结果整理加工后得到的。“词频统计”是国家科委下达、并委托国家标准局主管的科研项目。主办单位、参加单位以及主要参加人员有：

主办单位　北京航空航天大学

刘源（负责人），梁南元（后期负责人），孙伟，

杨铁鹰，王德进，张社英，马广申，王以德等

参加单位

中国人民大学	郑虎，黄建民等
北京大学	李一华等
吉林大学	管纪文，王锡龙，谷新英，王树义等
北京师范大学	何克抗等
武汉大学	张普等
四川师范大学	肖启洪等
南开大学	李约瑟等
原航天部二院	蒋赞平，祝楚恒等
新华社技术研究所	朱坤等

为了保证这个项目高质量完成，国家标准局特聘请了 16 位专家担任本项目的顾问。他们是：

钱伟长	学部委员，中文信息学会理事长，中国政协副主席
王力	学部委员，北京大学教授
王湘浩	学部委员，吉林大学副校长
支秉彝	学部委员，中文信息学会副理事长
陈力为	教授，中文信息学会副理事长
周有光	研究员，国家语委会
杜松寿	研究员，国家语委会
刘涌泉	研究员，中文信息学会副理事长
张其璿	教授，云南大学
许孔时	研究员，中科院软件所所长，中文信息学会副理事长
郑易里	高级编辑，中国农科院
唐亚伟	教授，北京速记协会理事长
王宗柏	中国人民大学语言所所长
欧阳文道	中文信息学会理事，中科院干部管理学院
李金铠	教授，中文信息学会常务理事
扶良文	高工，中文信息学会理事

经过北京航空航天大学和全体参加单位 4 年多的努力，“词频统计”于 1986 年 5 月结束，于 1986 年 6 月通过了由国家科委批准、并委托国家标准局主持的国家级鉴定。这个项目共计：

投入人力：	103.5 人年；
使用终端机时：	18490 小时；
使用 CPU 时间：	1152 小时；
选材：	3 亿汉字；
抽样：	约 2500 万汉字；
编码输入：	2000 余万汉字；

对 1919 年至 1982 年（分为 4 个时期）、社会科学和自然科学（分为 10 类子学科）的汉语语

言材料，分时期、分学科进行了词频统计。

这次词频统计是国内进行的一次规模最大、被统计材料时间分布最长、分科最多的汉语词频统计。据了解，国外还未进行过这样规模巨大的汉语词频统计工作。它具有选材抽样分布合理、背景干扰小、分词标准一致、统计精度较高等优点。

在进行统计工作的同时，还研制成功了国内第一个完整的汉语词频统计软件系统（国外未见过关于此类系统的报道）。其中的计算机汉语分词系统为国际首创，建立了一个有130691个词条的计算机词典，实现了一个具有诸多属性的汉字信息库，可为语言文字和信息处理工作者提供大量科研数据信息。

在词频统计的同时，还进行了字频统计和14种标点符号的统计。有关汉字频度统计结果将在近期出版。14种标点符号的统计频度见表3。

表3. 标点符号统计频度表（双引号、圆括号、书名号左右部分都计数）

标点符号	逗号	句号	分号	顿号	问号	惊叹号	冒号
统计频度	938914	457396	33316	156876	21455	12470	54177
统计频率	46.04%	22.43%	1.63%	7.69%	1.05%	0.61%	2.66%
标点符号	破折号	省略号	双引号	圆括号	书名号	间隔号	连接号
统计频度	10436	33749	164212	124842	25087	5381	932
统计频率	0.51%	1.66%	8.05%	6.12%	1.23%	0.27%	0.05%

汉语词频统计不仅是一项工作量巨大的工程，它也涉及到统计学、汉语语言理论和计算机技术等研究领域的许多问题，它与抽样理论、汉语词的定义、词的划分、知识表达、语法自动分析、语义理解等有直接关系。

“词频统计”的研制流程如下：

1. 选材抽样；
2. 编码输入；
3. 自动分词、统计；
4. 打印输出。

以下按照研制流程介绍本项目的研制情况。

3.1 选材抽样

原始母体选材量为3亿汉字，然后从中抽样约2500万汉字。先选材后抽样具有背景干扰小、统计结果可信度高的优点。

按照“词频统计”的目的，定义原始母体的时间范围是1919年至1982年，选材时又将其划分为4个时期：

第一时期：1919年至1949年（民主革命时期）；

第二时期：1950年至1965年（社会主义建设时期）；

第三时期：1966年至1976年（文革时期）；

第四时期：1977年至1982年（调整改革时期）。

各时期所占选材比例，是按照厚今薄古的原则分配的。文革时期因受影响严重，故所占比例较小。详细情况见表4。

表 4.“现代汉语词频统计”选材抽样明细表

科别 / 时期	社会科学(S)						自然科学(N)					
	1 文体生活	2 历史哲学	3 政治经济	4 新闻报导	5 文学艺术	社会科学总和	1 建筑运输	2 农林牧渔	3 电子轻化工	4 重工业	5 基础科学	自然科学总和
第一时期 1919～1949	0.54	1.08	1.61	1.62	2.15	7						
第二时期 1950～1966	1.08	2.16	3.22	3.24	4.3	14	1.0	2.0	2.0	2.0	5.0	12
第三时期 1967～1976	0.756	1.512	2.254	2.268	3.01	9.8	0.5	1.0	1.0	1.0	2.5	6
第四时期 1977～1981	3.024	6.048	9.016	9.072	12.04	39.2	1.0	2.0	2.0	2.0	5.0	12
时期总和 1919～1981	5.4	10.8	16.1	16.2	21.5	70	2.5	5.0	5.0	5.0	12.5	30

每个时期的选材分为社会科学和自然科学两类。社会科学又分为 5 个子类：文体生活（包括服装、食谱、旅游、集邮等），历史、哲学（包括心理学、教育学、美学、社会学等），政治、经济（包括财贸、统计、管理等），新闻报导，文学、艺术（包括小说、散文、说唱文学等）。自然科学也分为 5 个子类：建筑、运输（包括邮政），农、林、牧、渔，轻工业（包括电子、日用化工、食品、纺织等），重工业（包括矿山、冶金、机械、能源等），基础科学（包括数、理、化、生、天、地）。各子类所占比例是按照通用频度统计的要求分配的。较为大众化的学科比例较大，反之则较小。社会科学和自然科学的比例为 7：3。第一时期自然科学类未选材，这是因为在这一时期内中国的自然科学著作很少。

选材来源有：报纸期刊，教材，通俗读物，科普读物等。选材不能是翻译作品，而应是有代表性的名作家的作品以及其他语言规范化的作品。

选材结束后，从母体中抽样约 2500 万，形成统计样本。采用了随机和有规律（等距、分层等）相结合的抽样方法。

选材及抽样方案经过项目顾问委员会 16 位顾问的审定。

3.2 编码输入

汉字编码采用了北京航空航天大学研制的“词字混合码”汉字编码方案。这个方案的特点是纯拼音、可区分多音字。由于可区别多音字，不仅提高了自动分词精度，而且使统计结果更精确，可为更多的领域使用。

3.3 自动分词、统计

“词频统计”首次使用计算机自动分词和频度统计。自动分词的优点是速度快、分词形式统一、便于复查和重新统计。北京航空航天大学在国内外第一次科学地研究了自动分词的可行性，在大量统计的基础上，证明了自动分词是可行的。“词频统计”实现的第一个汉语自动分词系统使用知识库时，分词精度约每 500 个字发生一次错误切分。因为时间和经费等原因，“词频统计”自动分词时没有使用知识库。经抽样检查，自动分词的精度为每 180 个字发生一次错误切分。

分词词典的收词原则就是词的切分规则。由于“词频统计”是一次通用词频统计，为了能使统计结果可以为更广泛的领域所使用，分词词典的收词范围较宽。不同的用户可以根据自己的需要分解不需要的词条，将其频度迭加到相应的词条上。分词词典的词条来自：

1.《现代汉语词典》；
2.《辞海》；
3.《汉英词典》；
4.《汉法词典》；
5.《汉日词典》；
6.《标准汉英词典》；
7.《常用字构词词典》；
8.《列车时刻表》中全国火车站名；
9.《行政区划表》中地区以上的地名；
10.《国际著名人名录》中的人名；
11.《现代汉语八百词》；
12.《中国地名手册》；
13.《世界地名手册》；
14.《汉语拼音词汇》；
15.《汉语小词典》；
16.《成语词典》；
17.《汉英小词典》；
18.《常用词语三用词典》；
19.《常用汉字音形教学手册》；
20.《外国哲学社会科学人名录》中的人名；
21.《当代国际人物词典》中的人名；
22.《世界报刊通讯社、电台译名手册》；
23.《现代汉语词表》。

经去重和整理，以及增加了一些新生词条，例如“五讲四美、优生”等，分词词典有词条130691，最长词条有7个字。

统计是分时期分科进行的，这样各时期各类子学科的统计结果可以应用于不同的领域。

3.4 打印输出

为满足各种查询输出要求，在实现词频统计软件系统时专门设计了一个查询输出子系统。用户可以通过一个查询语言，方便的选择需要输出的字、词、范围、频序、音序、升序、降序、每页栏数等。

四 鸣 谢

在“词频统计”研制过程中，课题组得到了国家科委、国家标准局、电子工业部计算机局、十六位顾问，以及国家语言文字工作委员会、中国中文信息学会和原航空工业部的指导和帮助；也得到了总参三部八局计算机处、上海中华信息技术总公司、中科院计算所、中国计算机服务公司、华夏公司的协助和支持，特别是北京海淀区京航科技服务公司和中国人民大学计算中心给予了财政支持和使用计算机的方便。在这里必须列出对此项目给予大力支持和帮助的人士，他们有傅洪畴、姚世全、吴青、戴荷生、金光、洪用对、汤丙午、宁金源、刘凤翘、郭景春、郝春民、

钟锡昌、刘菊芬、沈元等。对本项目给予直接指导的还有陈原、陈章太、何耀坤、廖幼鸣、陈昊苏、慈云桂、姜学锦、吴几康、傅永和、萨师煊、陈树楷、李峰、唐亚伟、吕必松、竺乃刚、石云程、褚善元、张淞芝、吴克忠、殷志鹤等。借此机会，对以上单位和专家以及所有支持过本项目工作的单位和专家表示衷心的感谢。

本词典在编纂整理、出版过程中，得到了国家标准局、宇航出版社的指导和帮助。国家政协副主席、中国中文信息学会理事长、本项目顾问钱伟长教授欣然为本词典写了前言。姚世全、郝芬、杨德、赵珀璋、赵艳华等同志在整理过程中审阅了全部词条，提出了很多宝贵意见。本词典的责任编辑廖寿琪同志在出版过程中给予了不断的指导和支持。为了使国外读者能直接了解本词典，特请刘荣科、窦振波、姜育君、何妙生、范娟娟等五同志，将钱伟长教授写的前言分别译成了英、日、法、德、俄五种外国文字。我们谨对以上单位和专家表示衷心的感谢。

如果本词典在选词、体例、注音等方面存在错误，这些错误是在整理编纂过程中导致的。我们热切的欢迎读者的批评、建议和意见。

编者　刘源　梁南元　王德进　张社英　杨铁鹰　揭春雨　孙伟

1988 年 5 月

目　　录