中国语言生活绿皮书
国家语言文字工作委员会发布

YW B001

中国 Language Situation in China:2005

语言生活状况报告

下编

国家语言资源监测与研究中心 编

ZHONGGUO YUYAN SHENGHUO ZHUANGKUANG BAOGAO (2005)

2005

商務印書館 THE COMMERCIAL PRESS

中国语言生活绿皮书

国家语言文字工作委员会发布

中国语言生活状况报告

(2005)

下　编

国家语言资源监测与研究中心　编

商務印書館

2006 · 北京

图书在版编目（CIP）数据

中国语言生活状况报告. 2005. 下编/国家语言资源监测与研究中心编. —北京：商务印书馆，2006
ISBN 7-100-05228-9

Ⅰ. 中… Ⅱ. 国… Ⅲ. 社会语言学—研究报告—中国—2005 Ⅳ. H1

中国版本图书馆 CIP 数据核字(2006)第 109771 号

ZHŌNGGUÓ YǓYÁN SHĒNGHUÓ ZHUÀNGKUÀNG BÀOGÀO (2005)

中国语言生活状况报告（2005）

下　编

国家语言资源监测与研究中心　编

商务印书馆出版
（北京王府井大街36号　邮政编码 100710）
商务印书馆发行
北京瑞古冠中印刷厂印刷
ISBN 7-100-05228-9/H·1263

2006 年 9 月第 1 版　　开本 787×1092　1/16
2006 年 9 月北京第 1 次印刷　　印张 33¼
印数 5 000 册

定价：53.00 元

顾　问　许嘉璐　赵沁平
策　划　教育部语言文字信息管理司
主　编　李宇明
审　订　陈章太　戴庆厦　陆俭明　邢福义

上　编

主　编　周庆生
副主编　郭　熙　周洪波
作　者　（按音序排列）

陈　敏　戴红亮　丁石庆　冯学锋　高建平
郭　熙　黄晓蕾　黄　翊　汲传波　江　获
李晟宇　李晓华　李旭练　李艳华　刘海涛
刘宏宇　刘　青　娄开阳　吕　禾　牟云锋
司红霞　汪　磊　王丹卉　王　晖　王培光
王铁琨　魏　丹　谢俊英　徐世璇　于　虹
袁钟瑞　詹卫东　张　黎　郑梦娟　周洪波
周庆生　邹海清　邹玉华

下　编

主　编　王铁琨
副主编　张　普
作　者　（按音序排列）

安　娜　韩秀娟　何婷婷　侯　敏　李晋霞
李青梅　刘一玲　刘　云　秦　鹏　瞿国忠
史艳岚　苏新春　滕永林　涂新辉　汪　磊
王　彬　王铁琨　王依然　魏　励　文采菊
杨尔弘　余桂林　张　普　张书岩　张小鹏
张　勇　周　鑫　邹红建

目　　录

第一部分　调查报告

报纸、广播电视、网络用字用词调查 …… 003
高校网络媒体 BBS 用字用语调查 …… 017
中国报纸十大流行语 …… 022

第二部分　调查数据

表 1　报纸、广播电视、网络用字总表 …… 033
表 2　2500 高频字与一级常用字比较 …… 253
表 3　3500 高频字与《现代汉语常用字表》比较 …… 254
表 4　前 7000 字与《现代汉语通用字表》比较 …… 255
表 5　“用字总表”中未出现的通用字 …… 256
表 6　繁体字 …… 257
表 7　异体字 …… 262
表 8　不合现行规范的类推简化字 …… 265
表 9　旧印刷字形 …… 266
表 10　旧计量单位用字 …… 267
表 11　方言字 …… 268
表 12　日本汉字 …… 269
表 13　报纸、广播电视、网络高频词语表 …… 270

表 14　报纸、广播电视、网络高频词语用字表 …… 465
表 15　标点 …… 487
表 16　数字 …… 491
表 17　字母 …… 494
表 18　其他符号 …… 500
表 19　2005 年高校网络媒体 BBS 用语表 …… 504

后记 …… 526

第一部分

调查报告

报纸、广播电视、网络用字用词调查

语言文字是人类最重要的信息载体和交际工具，是“软国力”的重要组成部分。语言生活是社会生活的重要内容，随着社会的发展变化，语言生活也在不断地发展变化。为了切实掌握我国当前语言国情状况，及时把握我国年度用语用字的第一手资料，为国家语言政策的调整和制定以及语言文字规范标准的制定、修订提供参考，国家语言资源监测与研究中心对2005年报纸、广播电视、网络等媒体的汉字、词语、标点与符号的使用情况进行了调查，并就调查结果进行了初步分析。

一　调查使用的语料说明

调查语料分为平面媒体、有声媒体、网络媒体三种，共计892 034个文本文件，909 429 700字符次[①]，其中汉字出现732 143 010字次[②]。

（一）报纸

平面媒体选择了2005年15种报纸的网络版作为调查语料，选择时综合考虑了“发行量、发行地域、发行周期、媒体价值、阅读率”等五个因素。发行量主要根据2004年6月1日在伊斯坦布尔闭幕的第57届世界报业大会发布的最新“世界日报发行量前100名名单”（中国部分）；媒体价值主要根据世界品牌大会公布的2004年《中国500最具价值品牌》排行榜。

这15种报纸是（按音序排列）：《北京青年报》、《北京日报》、《北京晚报》、《法制日报》、《光明日报》、《广州日报》、《华西都市报》、《环球时报》、《今晚报》、《南方周末》、《人民日报》、《深圳特区报》、《羊城晚报》、《扬子晚报》、《中国青年报》。

---

① 字符次，指汉字以及字母、数字、标点、符号等在语料中出现的次数。

② 字次，指汉字在语料中出现的次数。

报纸语料共计 591 315 个文本，538 752 953 字符次，其中汉字出现 425 789 961 字次。

（二）广播电视

有声媒体选择了 2005 年广播电视一些节目文本作为调查语料。这些语料是电台、电视台播出的有声节目的录音或录像转写成的文本资料。选择时综合考虑了“传播媒介（广播、电视）、媒体级别（中央、地方）、播出时间（黄金时间、非黄金时间）、节目样态（独白、对话、综合）、文本现存（是否有转写好的文本）”五个因素。语料来源如下：

电视节目类：中央电视台、北京电视台、上海电视台、上海东方电视台、广东电视台、天津电视台、安徽电视台、山东电视台、长沙电视台、重庆电视台、东方卫视、广州电视台等 13 家电视台的 87 个栏目的 5 567 个节目的文本文件，共计 25 441 066字符次。

广播节目类：中央人民广播电台、北京人民广播电台、海峡之声广播电台、深圳广播电台、广东人民广播电台、天津人民广播电台、上海东方广播电台和中山广播电视台等 8 家广播电台的 57 个栏目的 2 337 个节目的文本文件，共计 5 753 846 字符次。

上述两类节目语料中，汉字共出现 25 845 303 字次，其中无法显示的字符 32 个，共计出现 36 字次。

（三）网络

网络媒体选择了新华网、人民网、中华网、中国新闻网、新浪网、网易等网站的新闻文本作为调查语料，共计 295 152 个文本，339 482 177 字符次，其中汉字出现 280 507 746 字次。

二 调查内容

本次的调查对象是汉字、词语、标点、符号等。汉字、词语的调查项目有“频次、频率、累加频率、文本数”等；标点、符号等的调查项目有“频次、文本数、所占比例”等。上述调查项目的含义及计算方法是：

频次：每一调查对象在所有语料中出现的次数。

文本数：语料中包含该调查对象的文本个数。

频率：每一调查对象的频次与整个语料所含调查对象总次数的比值，即：

$$F_i = n_i / N \times 100\%$$

其中：F_i 为调查对象 i 的频率，n_i 为调查对象 i 的出现次数，N 为语料中调查对象出现的总次数。

累加频率：所有调查对象按照频次降序排列，每一调查对象的频次同其前调查对象频次的累加和，与所有语料中调查对象总次数的比值，即：

$$A_i = \sum_{k=1}^{i} n_k / N \times 100\ \%$$

其中：A_i 为调查对象 i 的累加频率，n_k 为调查对象 k 的出现次数，N 为所有语料中调查对象的总次数。

任一标点、符号等占全部标点、符号的比例：每一标点、符号的频次与语料中标点、符号出现的总次数的百分比，即：

$$P_i = pn_i / Tp \times 100\ \%$$

其中：P_i 表示标点、符号 i 占整个标点、符号的比例，pn_i 为标点、符号 i 的频次，Tp 为语料中标点、符号出现的总次数。

任一标点、符号等占语料的比例：每一标点、符号的频次与语料总量的百分比，即：

$$CP_i = pn_i / Tc_1 \times 100\ \%$$

其中：CP_i 表示标点、符号 i 占整个语料的比例，pn_i 为标点、符号 i 的频次，Tc_1 为所考察语料中分词单位[①]的总次数或总字符数[②]。

标点是按照分词单位提取的，计算语料总量时，按照分词单位的出现次数计算。符号是按照单个字符提取的，计算语料总量时，按照总字符数计算。

三　调查结果

（一）汉字使用情况调查

说明：

1. 报纸文本是从网络下载的，没有与纸质版本作比较。广播电视语料也是

① 分词单位的含义见 010 页词语使用情况调查部分。

② 总字符数的含义见 006 页汉字使用情况调查部分。

从网络下载的，通过比较，发现与实际有声语料之间存在或大或小的差异。网络媒体的语料均来自各网站2005年创建的页面。上述三种语料均做了去除HTML标签信息和广告信息的处理。

2. 本次统计没有甄别文本中的别字。

3. 本次统计不包括以下两种字符：

(1) 汉字部件。共有25个，计529字次，主要出现在报纸语料和网络语料中。包含以下两类情况：

① 讲解汉字时用到的偏旁部首，如“言语的‘语’这个字旁边是个‘讠’字旁”。这种部件共出现四个：亻、辶、宀、讠；

② 拼字，大部分出现于人名、地名，如“讲述人：刘亻思亻思，14岁，树德实验中学”、“本市宝坻区林亭口镇帐房瞿阝村农民”；“广东中山南(艹朗)(上下结构)镇横门港码头彩旗飘扬”、“20岁的广西姑娘小(崩刂)昨天回广州了”。

(2) 乱码和无法显示的字符。这些字符共出现765个，计23 221字符次，占整个语料字符数的0.0026%。这些字符可以分为以下几类：

① 不同的报纸媒体因使用不同的排版系统，在进行版式转换时产生的特殊符号；

② 用来代表网页图片的符号；

③ 由于字符错位造成的符号。

调查结果：

1. 总字符数：指全部语料中汉字、标点、符号等的总量(不包括乱码、无法显示的字符和汉字部件)，计909 429 700字符次。

2. 字符种数：8 713个。这里的字符种，指不同形式的字符(包括汉字、标点、符号)。

3. 总汉字数：指全部语料中汉字出现的总字次，计732 143 010字次。

4. 字种数：8 128个。这里的字种，指字形不同的汉字。

报纸、广播电视、网络用字总表，见033页表1。

5. 共用字种数：5 606个。这里的共用字种 指报纸、广播电视、网络都用到的汉字。

6. 独用字种数：见007页表1-1。这里的独用字种，指本调查中只在报纸、广播电视、网络某一媒体中出现的汉字。

表 1-1　汉字使用情况

媒体	总字次	字种数	共用字种数	独用字种数
报纸	425 789 961	8 038	5 606	1 628
广播电视	25 845 303	5 761	5 606	45
网络	280 507 746	6 351	5 606	39
总计	732 143 010	8 128	5 606	

各媒体共用、独用字具体数据见 033 页表 1《报纸、广播电视、网络用字总表》中“共用独用”栏，其中 A 代表报纸，B 代表广播电视，C 代表网络。如果一个字在该栏标有 ABC，表示该字是三种媒体共用字；如果标 AB，则是报纸、广播电视共用字；标 BC，则是广播电视、网络共用字；标 AC，则是报纸、网络共用字；只标有一个字母的，则表示是该媒体的独用字。

数据表明，报纸作为平面媒体，用字量大，独用字最多，显示出较浓重的书面语特点。广播电视用字量较少，这和它是有声语言，以口耳相传为主有关。

7. 汉字的覆盖率。

表 1-2　汉字对语料的覆盖情况

覆盖率	达到 80% 的字种数	达到 90% 的字种数	达到 99% 的字种数
全部语料	581	934	2 314
报纸	585	937	2 345
广播电视	507	869	2 303
网络	557	897	2 214

从表 1-2 可以看出，581 个汉字就覆盖全部语料的 80%，934 个汉字就覆盖全部语料的 90%。这从一个侧面说明学习汉语并不像有些人想象的那么难，掌握了 581 个汉字，差不多就可以认识媒体上 80%的文字。但这并不意味着掌握了 581 个汉字就能读懂媒体上 80%的内容，因为语言理解、语言学习还涉及到掌握词汇量的多少和文化背景等诸多因素。

8. 2500 高频字与一级常用字的比较。

本“用字总表”(8 128 字)前 2500 高频字与《现代汉语常用字表》(国家语言文字工作委员会、国家教育委员会 1988 年联合发布)一级常用字(2500 字)比较，其结果是，《现代汉语常用字表》“一级常用字”中有 357 字在本“用字总表”前 2500 高频字中没有出现。见 253 页表 2。

9. 3500 高频字与《现代汉语常用字表》的比较。

本“用字总表”(8 128 字)前 3500 高频字与《现代汉语常用字表》(国家语言文字工作委员会、国家教育委员会 1988 年联合发布)3500 字比较，其结果是，《现代汉语

常用字表》中有398字在本"用字总表"前3500高频字中没有出现。见254页表3。"用字总表"(8 128字)中包含了《现代汉语常用字表》的全部汉字。

10. 前7000字与《现代汉语通用字表》的比较。

本"用字总表"(8 128字)前7000字与《现代汉语通用字表》(国家语言文字工作委员会、中华人民共和国新闻出版署1988年联合发布)7000字比较,其结果是,《现代汉语通用字表》中有506字在本"用字总表"前7000字中没有出现。见255页表4。

11. "用字总表"与《现代汉语通用字表》的比较。

调查结果显示,《现代汉语通用字表》中有244个通用字在本"用字总表"(8 128字)中没有出现。见256页表5。

从这次调查所得的"用字总表"与1988年发布的《现代汉语常用字表》、《现代汉语通用字表》的比较中可以看出,高频汉字的使用发生了某些变化,这里可能有语料领域不同的原因,但它在一定程度上折射出当今社会生活的变化以及人们表达焦点的变化。如《现代汉语常用字表》前2500字中与农业生活有关的"驴、骡、锄、铲、犁、镰、秧、秆、茎、稼、禾、箩、筐、粪、浇"等字从高频字中退出来,似乎可以说明随着社会的发展,传统的农业模式离我们慢慢远去,我国正在向着工业化、现代化迈进;过去与日常生活密切相关的"绸、缎、绢、袄、袍、饺、馒、薯、糠"从高频字中淡出,则反映出人们生存方式、生活方式的某些改变;"舅、姨、婶、侄"等亲属称谓字的使用率减少,可能与多年来执行的计划生育政策有关;而"鸽、狐、狸、蚂、蚁、龟、蝴、蜻、蜓、蛙、鹊、雁、骆、驼、葵、椒、蕉、柿、橡、棕、榆、芦、桐"等在"用字总表"高频字中没有出现,是不是意味着在迈向工业化、现代化的同时,人们也在远离自然?像"鄙、傻、愚、蠢、笨、懒、惰、怠"这些带有贬义倾向的字眼使用频度下降,也许可以从一个侧面说明社会文明程度的提高,人们在对他人进行评价时会尽量少用这类带有刺激性的字眼。(当然,"用字总表"中也有贬义字,如"淫、诡、诈、歹、尴、尬"等。)而"用字总表"中出现较多的姓名用字,如"邓、姚、蔡、邢、侯、郭、坤、翔、婷、菲、琛、鑫、娜"等,这大约与这次调查选用的新闻语料较多有关。

12. 报纸、广播电视、网络高频词语用字统计。

报纸、广播电视、网络高频词语10 356条,见270页表13;共使用汉字2 463个,见465页表14。

13. 汉字的其他情况。

对“用字总表”(8 128 字)中的繁体字、异体字、不合现行规范的类推简化字、旧印刷字形、旧计量单位用字、方言字、日本汉字等七类字,分别用专门符号作了标示,详见 257—269 表 6—表 12。总体情况见 009 页表 1-3 所示。

表 1-3 汉字的其他情况统计

	繁体字	异体字	不合现行规范的类推简化字	旧印刷字形	旧计量单位用字	方言字	日本汉字
全部语料	361	193	7	47	3	36	62
报纸	339	188	7	47	3	36	61
广播电视	33	18	4	5	0	2	2
网络	2	11	5	0	0	4	0

繁体字使用主要有五种情况:

(1) 讲解汉字时使用了繁体字,如“只有心在中间时才生‘爰’(繁体为‘愛’)”。

(2) 人名、地名中使用了繁体字,如:“谭文鋕”、“吴矇矇”、“鬱陵島”。

(3) 引用文言和古书名称时使用了繁体字,如“君之视臣如土芥,则臣视君如寇讎”、“明代万历年间刻本的《殊域周咨録》”。

(4) 不小心录入了繁体字,有些繁体字在文本中只是偶尔出现,这可能是由于录入人员在选字时误选了繁体字。如一篇文章中“一辈子”出现了五次,只有一次写成“輩”,其余的都是“辈”。

(5) 有意使用繁体字,表现为一段文字中出现多个繁体字,如“東京高等法院最終認定文部省對家永三郎教科書上‘南京大屠殺’和‘七三一部隊’等 4 個問題的修改違法”、“最佳當代節奏布魯斯 RB 專輯”、“最佳鄉村樂隊”。但语料中这种情况极少。

异体字使用主要有三种情况:

(1) 讲解汉字时使用了异体字,如“文史字本索引以阮元编辑的《十三经注疏》为底本。……因此在严格控制下,对部分异体字,像‘群’和‘羣’、‘恆’和‘恒’、‘昏’和‘昏’、‘賛’和‘贊’、‘嚐’和‘嘗’、‘鶏’和‘雞’、‘厩’和‘廄’、‘厨’和‘廚’、‘柰’和‘奈’、‘却’和‘卻’等字进行了必要的统一。”

(2) 人名、地名中使用了异体字,如:“龙淼”、“高喆”、“王堃”、“黄霑”、“迺兹府”、“高砦村”。

(3) 个别的误用,如:“眯”写成“瞇”,“藤”写成“籐”。

从上述情况来看,讲解汉字时涉及繁体字和异体字以及文言引用中使用繁体字均属社会用字的正常现象;繁体字和异体字主要出现在人名和地名中;繁字的(4)、(5)和异体字的(3)虽然出现频率不高,但应引起注意,这和某些汉字输

入法提示备选框中繁体字、异体字与简体字并存有关，也与网络用字缺少必要的管理措施和有力的把关人员有关。

方言字使用主要有两种情况：

(1) 某些地名用字，如两个使用频率最高的方言字“碰”“垴”都是地名用字：“骆驼碰子、关家垴”。

(2) 某些方言区的报纸网络版上出现的方言字。

至于日本汉字、旧计量单位用字、旧印刷字形、不合现行规范的类推简化字等，在语料中出现的频率都非常低，除个别人名、地名用字外，大多数和汉字输入法提示备选框中各种字形并存、录入人员误选有关，并非有意为之。

此外，“用字总表”中还有 7 个古汉字：“糸、冂、隹、尢、厶、釆、丌”。它们主要出现在网络和报纸网络版语料中，频率极低，除“厶”为人名用字外，大部分属于误用了别字。如“糸”(mì，又读 sī)是“系”的误写，“冂”(wéi，又读 guó)是“口”的误写，“釆”(biàn)是“采”的误写，“尢”(wāng，又读 yóu)是“尤”的误写。分析起来，这种现象的出现也与汉字输入法提示备选框中各种字形并存、录入人员误选有关。

(二)词语使用情况调查

说明：

1. 本次调查使用的分词软件是中国科学院自动化研究所研制的分词标注系统，该系统在 2004 年度“863 中文信息处理与智能人机接口评测”中，取得了较好成绩。[①]

2. 歧义切分、未登录词识别等，目前仍然是计算机自动分词无法完全解决的问题。这些问题会对词语统计和分析结果带来一定影响。不过，由于所统计的语料数量巨大，同时又对切分后的高频切分单位进行了排查处理，可以在一定程度上使上述问题对统计结果的影响降到最低程度。

3. 本次调查不考虑同一词形不同词性或词义的区别。

调查结果：

1. 分词单位总数：指由分词软件对语料切分得到的字符串的总数，计 489 240 995词次[②]。

① 参见 www.863.org.

② 词次，在不同的上下文中，既可以指由分词软件对语料切分得到的所有字符串即分词单位出现的次数，也可以指分词单位中减去标点、符号、纯西文分词单位后的分词单位出现的次数。

2. 总词语数：指分词单位总数减去标点、符号、纯西文分词单位后的分词单位数量，计 416 090 995 词次。

3. 词种数：1 651 749 个。这里的词种，指去除标点、符号、纯西文分词单位后的不同形式的分词单位。

4. 共用词种数：106 111 个。这里的共用词种，指报纸、广播电视、网络等媒体都用到的词语。

词语使用的总体情况如本页表 1-4 所示。

表 1-4　词语使用情况

媒体	总词语数	词种数	共用词种数
报纸	234 475 341	1 132 165	106 111
广播电视	15 889 152	161 724	106 111
网络	165 726 502	847 971	106 111
总计	416 090 995	1 651 749	106 111

全部语料的词种数超过 165 万，其中人名约 61 万条，地名约 24 万条，机构名约 59 万条，时间词语约 10 万条，分别占总数的 36.9%、14.5%、35.7%、6%。其他词种数 11 万条，约占总数的 7%，其中大部分是语文词。各类词语的分布如本页图 1-1 所示。

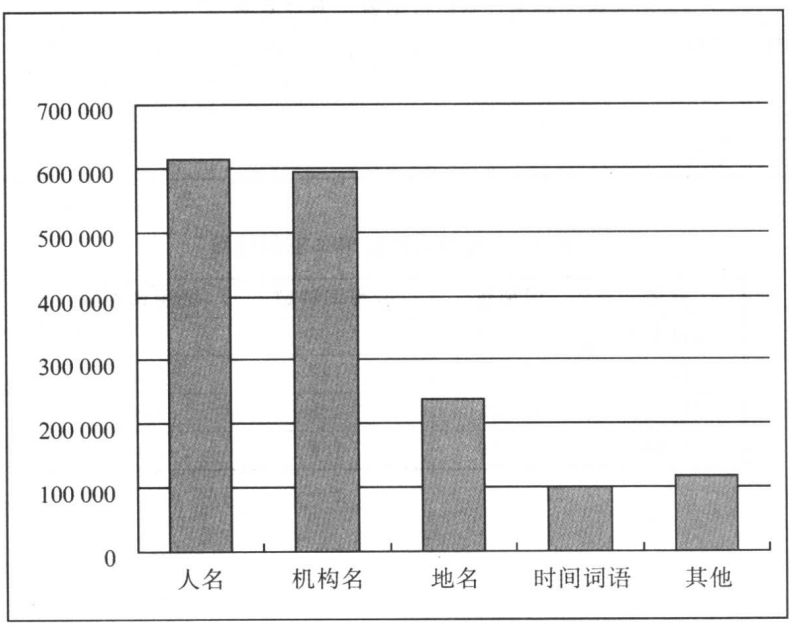

图 1-1　各类词语的词种数