报纸、广播电视、网络(新闻)用字用语调查

为实现语言文字跟踪调查的连续性,在2005年报纸、广播电视、网络(新闻)用字用语调查的基础上,国家语言资源监测与研究中心利用国家语言资源监测语料库(包括平面媒体、有声媒体、网络媒体)对2006年报纸、广播电视、网络(新闻)用字用语的情况进行了调查。

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1 311 749个文本文件,1 170 367 879字符次(包括标点、符号及西文字母、数字等,下同),其中汉字出现978 994 406字次。

(一)报纸

平面媒体选择了2006年15种报纸作为调查语料,报纸选择时综合考虑了"发行量、发行地域、发行周期、媒体价值、阅读率"五种因素。发行量参考了2005年5月31日在韩国汉城(首尔)举行的第58届世界报业大会发布的"2005年世界日报发行量前100名排行榜"(中国部分);媒体价值参考了2005年8月6日在北京举行的世界品牌大会发布的2005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》、《经济日报》、《南方周末》、《钱江晚报》、《人民日报》、《深圳特区报》、《中国青年报》。

报纸语料共计648 607个文本,471 860 752字符次,其中汉字出现399 488 842字次。

(二)广播电视

广播电视语料是电台或电视台播出的录音或录像的文本转写资料。语料选

择的主要依据是流通度(即节目收视率),涉及因素包括:传播媒介(广播、电视)、媒体级别(中央、地方)、传播广度(是否上星)、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)等。

2006 年选取的广播电视语料如下:

电视节目语料:中央电视台、中国教育电视台、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、南方电视台、山东电视台、湖北电视台等 10 家电视台 165 个栏目的 12 113 个节目的文本文件。

广播节目语料:中央人民广播电台、北京人民广播电台、广东人民广播电台、海峡之声广播电台、深圳人民广播电台、天津人民广播电台、山东人民广播电台等 7 家广播电台 51 个栏目的 1 600 个节目的文本文件。

广播电视语料总数为 64 408 209 字符次,其中汉字出现 53 029 167 字次。

(三)网络(新闻)

网络媒体采集了新浪、网易、腾讯、Tom、搜狐等 5 家网络门户网站的 2006 年全年的新闻语料,共计 649 429 个文本,634 098 918 字符次,其中汉字出现 526 476 397字次。

二　调查内容与方法

本次的调查对象是汉字和词语。调查项目主要有“频次、频率、累加频率、出现文本数、使用率、累加使用率”等。其中,频次、频率、累加频率、出现文本数的含义及计算方法同 2005 年的《报纸、广播电视、网络用字用词调查》[①]。“使用率”的计算方法有所改进。2005 年中“使用率”的计算公式为:

$$D_i = t_i / T \qquad (1)$$

$$U_i = F_i \times D_i \qquad (2)$$

其中:D_i 为 i 号字或词的分布率,t_i 为 i 号字或词的出现文本数,T 为所有语料的文本总数;U_i 为 i 号字或词的使用率,F_i 为 i 号字或词的频率。这样计算的结果,所有字或词的累加使用率小于 1。为了归一化,本次调查中把公式(2)改为(3):

① 王铁琨主编《中国语言生活状况报告(2005)》(下编)第 003—016 页,商务印书馆 2006 年版。

$$U_i = \frac{F_i \times D_i}{\sum_{j \in V} (F_j \times D_j)} \qquad (3)$$

其中分母为归一化项,V 表示所有字种或词种。这样累加使用率就可以归为 1 了。①

三　调查结果

(一)汉字使用情况调查

说明:

1. 报纸语料是网络版的。广播电视语料是由广播电视节目转写的文本,与原始有声语料之间存在某些差异。网络语料来自各网站 2006 年创建的页面。上述三种语料均作了去除 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. 本次统计不包括汉字部件、乱码以及无法显示的字符②。

4. 本次调查报纸语料规模与去年基本持平,广播电视和网络语料规模有所扩大,语料总量比 2005 年多出近 2 亿字。

调查结果:

1. 基本情况

(1)总字符数:指全部语料中汉字、标点、符号等的总量(不包括无法显示的字符),计 1 170 367 879 字符次。

(2)字符种数:10 781 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 978 994 406 字次。

(4)字种数:9 231 个。这里的字种,指字形不同的汉字。

(5)共用字种数:6 032 个。这里的共用字种,指报纸、广播电视、网络(新闻)三种媒体都用到的汉字。

---

① 侯敏《语言监测与词语的计量研究》,载《中文信息处理前沿进展》第 102 页,清华大学出版社 2006 年版。

② 关于乱码和无法显示的字符的具体说明,见王铁琨主编《中国语言生活状况报告(2005)》(下编)第 006 页,商务印书馆 2006 年版。

(6)部分共用字种数:1 367 个。这里的部分共用字种,指只在某两种媒体中出现的汉字。

(7)独用字种数:1 832 个。这里的独用字种,指只在报纸、广播电视、网络(新闻)某一媒体中出现的汉字。

汉字使用情况的具体数据见表 1-1。

表 1-1　2006 年汉字使用情况

媒　体	总字次	字种数	共用字种数	部分共用字种数	独用字种数
报　纸	399 488 842	8 326	6 032	1 317	977
广播电视	53 029 167	6 194	6 032	98	64
网络(新闻)	526 476 397	8 142	6 032	1 319	791
全部语料	978 994 406	9 231	6 032	1 367	1 832

2. 根据频率、使用率排序所得字表的比较

频率与使用率是两个不同又有关联的概念。使用率是在频率的基础上又加进了分布率的概念。在一定范围内,按频率还是按使用率排序来提取常用词,结果不完全一样。我们对 2006 年用字总表分别按频率和使用率进行了排序统计,表 1-2 显示了前 200、500、1 000、2 000、3 000 字二者的区别。

表 1-2　2006 年用字总表按频率、使用率排序比较

	相同字数	按频率排序独现字	按使用率排序独现字
前 200 字	186	赛球价斯女海社东交水管商么设(14 字)	由及受无向只果接强解常打任数(14 字)
前 500 字	475	农村校罗游销林牌户男病火防巴险江党购维星觉它职钱卡(25 字)	且终却负照必响效除功象往断曾半预另满言存历极致双类(25 字)
前 1 000 字	970	朗租萨券篮毒吴妈贸弹软刑姆藏瑞孙鲁牛患贷川赵津您姐课锦杰哥染(30 字)	异启努距综键穿享熟呼既怀询载秘端杂避盛误幸静娱玩辑延暴探寻释(30 字)
前 2 000 字	1 958	犬菌艇墓乙妮湘寺炭仲肝膜葡肇渔亨琼侨邵桶萄厨邱脂赂乒尿蹈詹坤兑丛狮郅肿募婷岳鹿帖挪饼(42 字)	饱呆衷浏框逢弥掀脆斥慕躲闷囊舆尬屈尴赋渴帽煌滞罕弯糕锻悠衔袖鼻爽擅聪撒眉恨皆砸嫁鼎飘(42 字)

续表

	相同字数	按频率排序独现字	按使用率排序独现字
前 3 000 字	2 947	蕊 猩 镐 汶 薯 鲸 筝 芹 鸽 驴 伽 娥 泸 舜 芸 钙 婕 敖 佟 橱 榆 咪 豚 俺 隋 氢 祁 蚁 琛 禄 佼 栗 葫 樟 雍 坍 泵 汀 窟 柬 阪 氯 棺 磷 茄 胚 氨 瑛 冥 墩 沐 缮 岐 (53 字)	踊 辗 俨 窘 捂 捍 抒 俯 蹄 煞 拽 吟 抨 拭 矢 挟 辄 伺 镶 蹦 辍 旷 咒 窥 萃 絮 敷 汹 黯 陡 炙 譬 糙 沮 爪 叭 藉 揪 靴 棕 荫 憋 匮 赐 拇 瞎 绚 蝉 聆 烁 拎 昼 窍 (53 字)

从表中可以看出,按频率统计,如“球、赛、水、海、农、村”等记录实词的字,由于论述主题的关系可能在一些文本中高频出现;但按使用率统计,由于它们出现的文本范围不广,就不得不让位于一些记载使用范围较广的功能词的字,如“由、及、向、只、且、却”等,这在前 500 字中表现比较明显。前 1 000、前 2 000 字段比较明显的是,按频率统计会有一些人名用字或表示特别事件的字进入统计范围,如“朗、瑞、姆、吴、孙、鲁、赵、郅、婷”等;而按使用率,则一些普通语词用字占统计优势,如“异、启、努、穿、享、熟、饱、呆”。频率和使用率都很重要,不同的研究目标会从不同的角度统计,因此我们在调查表中一般只给出原始数据——频次和出现文本数,目标数据将由研究者根据需要自行计算。2006 年汉字频率与使用率情况如图 1-1、图 1-2、图 1-3。

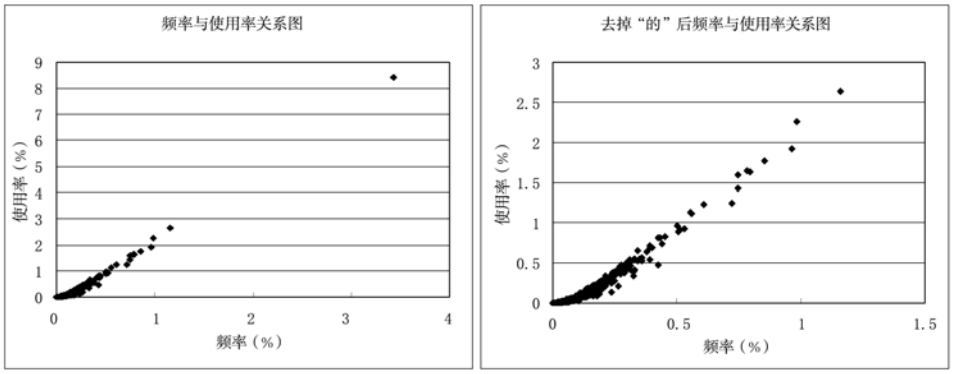

图 1-1　2006 年汉字频率与使用率关系

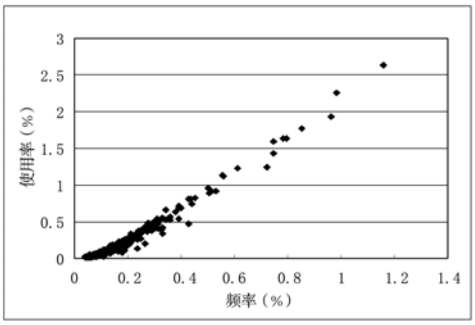

图 1-2　去掉“的”后覆盖率达 80％(591 字)时频率与使用率关系

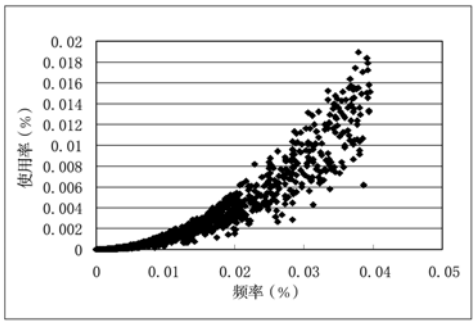

图 1-3　去掉“的”后覆盖率在 80％(592 字)—100％(9231 字)时频率与使用率关系

从图 1-1 的左图可以看出，汉字的频率和使用率基本为线性正相关，即频率高的，使用率也高，这是一个总的趋势。但从去掉最高频的“的”字，将坐标放大后的右图，可以看出二者之间只是大体上呈线性。两者的相关系数为0.955 2，这一数值既显示出较强的相关度，又说明二者还有不同。图 1-2 和图 1-3 是去掉“的”后覆盖率在前 80％和后 20％的汉字的频率和使用率关系图。图形表明，覆盖率在前 80％的字所表现出的趋势与图 1-1 所表现的大抵相似，而图 1-3 中，坐标被进一步放大，更加清楚地凸显出后 20％的汉字的频率和使用率之间，既有基本相关的一面，也有存在一定差异的一面。

3. 汉字的覆盖率

这里的覆盖率，是按汉字频率的累加来计算的。

表 1-3　2006 年汉字对语料的覆盖情况

覆盖率 / 语料	达到 80％的字种数	达到 90％的字种数	达到 99％的字种数	达到 100％的字种数
全部语料	591	958	2 377	9 231
报　纸	591	955	2 401	8 326
广播电视	528	904	2 379	6 194
网络(新闻)	589	954	2 340	8 142

4. 与现行字表的比较

(1) 2 500 高频字与一级常用字比较

“用字总表”(2006 年全部字种的集合，9 231 字)前 2 500 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(2 500 字)比较，其结果是，“用字总表”中有 331 字是“一级常用字”中

所没有的。这些字是：

尔 伊 媒 韩 圳 俄 频 诺 迪 萨 姆 措 综 伦 莱 辑 曼 菲 洛 澳 谐 姚 蒂 署
埃 杭 郭 谓 聘 账 凌 娜 屏 赫 艾 贾 霍 鹏 咨 沪 卢 弗 聊 邦 帕 翔 拟 冯 胎 拓
潘 邓 穆 颁 莞 韦 戈 曹 浦 敦 奈 耶 蔡 莉 兹 讼 氛 肖 彭 玛 秦 颖 徽 斌 琳 颇
玲 袁 芯 廷 履 逊 诈 蒋 茨 曝 吕 癌 枚 柯 吁 涵 厢 憾 汰 魏 怡 旭 硕 卓 崔 妆
谭 婴 莎 郁 庞 鸿 抑 妮 魅 侯 翰 湘 蕾 裸 仲 淀 涯 雇 晰 肇 辖 亨 歧 琼 邵 铭
邱 惟 赂 詹 坤 兑 郅 募 婷 岳 逻 契 宠 磊 砸 逸 楠 勃 舆 鼎 靖 邹 浏 晖 骚 靓
滞 雯 鲍 粤 娟 寓 衷 淑 坎 弥 尬 墅 尴 赋 罕 擅 碟 幽 铝 卿 薇 佐 擎 坞 啤 歹
绎 彬 廖 淫 蓉 嘛 勘 苑 琪 弘 坠 陌 郝 瓷 侠 菱 刹 豫 奢 尹 峻 瘤 镑 杉 骏 彦
轴 挫 瞬 呵 侣 熙 彰 撰 瑟 吻 爵 遂 晤 瘾 遏 赁 扳 揽 鑫 昊 淮 溢 冕 谍 崛 玮
珊 缉 薛 暨 萧 喻 奠 玫 槛 虐 殴 粹 澄 坪 轩 祭 咖 腕 函 瑜 逾 啡 睹 噪 昭 凸
睐 迄 殷 昔 瑰 啥 裔 莹 馨 澜 奎 醇 潭 渝 磋 逛 辐 仕 撼 腺 矶 铀 飙 霆 蟹 媳
巢 喀 寂 馈 讶 囚 蕴 钦 鹤 琦 俐 倪 懈 肪 炫 豹 儒 溃 晏 崩 埔 凰 芭 睿 瞩 腻
甸 衍 勋 俞 谴 阐 娅 婉 缔 窦 龚 匿 甫 沽 邢 赃 瘫 蔓 痴 藤 蔚

(2) 3 500 高频字与《现代汉语常用字表》比较

“用字总表”前 3 500 高频字与《现代汉语常用字表》3 500 字比较，其结果是，“用字总表”中有 388 字是《现代汉语常用字表》中所没有的。这些字是：

圳 迪 弗 莞 韦 耶 蔡 兹 斌 茨 曝 柯 怡 莎 妮 魅 肇 亨 邵 邱 惟 詹 郅 婷
磊 楠 邹 浏 晖 靓 雯 鲍 娟 尬 尴 佐 廖 嘛 苑 琪 弘 郝 尹 镑 彦 鑫 昊 崛 玮 缉
暨 槛 瑜 睐 裔 馨 奎 渝 磋 仕 矶 铀 飙 霆 喀 馈 琦 倪 晏 埔 睿 瞩 俞 娅 窦 龚
帷 淇 阜 麟 厄 倩 湛 媛 霖 瑶 禅 皓 哦 璇 弈 潇 韶 烨 禹 璐 炜 殡 滕 尧 绯 颐
茜 涅 禹 妃 驭 裴 牟 汕 鄂 珀 妍 侃 炳 丫 嫖 吾 钰 慑 孚 鲨 诠 咋 铉 嵩 翟 酋
斐 跻 圭 苯 阮 闫 茹 黛 迭 朔 卉 汶 甄 戛 冉 昕 籁 巅 曦 汝 赣 闵 盎 伽 娥 泸
舜 墟 渎 芸 婕 姬 敖 佟 郡 豚 隋 祁 岚 琛 禄 佼 亟 雍 坍 汀 阪 嘘 瑛 冥 缮 岐
岑 佘 邯 荫 镖 俨 辍 菁 辗 襄 哇 拽 娟 焱 孜 虞 萃 烯 釜 胺 漳 抨 狩 锂 飓 郴
翡 黏 姗 羚 羁 踝 绮 匡 俪 拎 铮 亵 禧 聆 辄 伎 煽 缪 黯 煲 曰 嘻 匮 炙 藉 骋
绚 髦 婧 钊 昱 邸 鳖 瑾 钛 峨 狄 峙 愣 忡 咎 郦 渲 郜 憬 贻 胤 睫 俑 梓 暧 佬
庚 瑕 娴 憧 瞿 韬 沁 铬 悖 枭 癫 祀 峪 荟 猝 渭 葆 氟 臻 袂 粽 钰 鹭 醛 毋 蹿
裘 茶 皖 骸 瀚 驿 峥 轶 榻 煜 抉 陀 兮 嘟 霄 铂 莓 蓓 狙 嘎 挝 宸 覃 潍 烽 跤
悸 蹊 陇 栩 亢 妞 邬 猥 恪 炯 祛 熠 飚 眩 厮 潼 咄 獗 汲 诧 媲 遛 疵 靳 萱 滇
汾 踵 酮 寰 喆 璋 骼 孰 妩 毗 笃 悚 茱 洱 祺 啪 喃 垣 沱 迥 酶 槟 沓 淼 洙 沂

琶 酯 恺 荃 惬 跆 釉 骐 簧 珂 铎 炖 咀 斡 甬 梵 矣 绥 郸 玺 獒 隽 嗯 戎 剽 姝

(3)前 7 000 字与《现代汉语通用字表》比较

“用字总表”前 7 000 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)7 000 字比较，其结果是，“用字总表”中有 517 字是《现代汉语通用字表》中所没有的。这些字是：

闫 飚 喆 淼 堃 濛 唰 堀 玘 晳 玟 碁 珺 仝 嫜 崚 珅 瘀 後 鲶 炘 吒 埗 镕 钜 昇 崧 捱 哒 弢 焗 狲 芃 槃 磡 迳 吋 祐 茺 锺 呎 贠 啫 內 迴 珮 抛 桠 戉 畊 市 纮 篠 芘 琍 孖 嘅 郈 诶 咁 垚 榎 徳 錸 囍 骱 樑 噘 锠 埭 甦 浬 崑 俣 劼 谘 郎 蚵 椝 潮 蹚 叡 艄 欹 玓 岜 铣 嫵 潡 铖 迺 椹 辻 龢 曌 艉 冇 稹 垴 腁 吔 係 噁 箹 步 旼 珪 倖 魟 咲 笹 佢 啲 猓 迺 峇 別 禛 惪 啰 朢 時 醣 晚 麼 霑 燚 炤 頫 囯 褫 珋 瑷 圖 飏 埇 鄌 睒 蒨 瑨 枬 畑 詠 尓 壘 鎔 褚 罣 勔 侣 珗 佳 吖 堦 咇 袴 啟 吶 剎 衹 絜 琣 誌 砦 氹 栻 鏸 呒 塝 岣 鉏 郚 锞 訢 搧 舥 砩 讃 畬 砬 鹏 瘰 來 炆 黃 翾 愛 暐 梶 韡 秝 鞔 峯 粿 呯 咗 採 腘 瓅 嚓 岍 岽 杩 癫 荳 沒 呑 睍 暸 孃 爀 甯 懍 伲 丟 荽 湑 給 蘋 穮 姮 江 佺 靚 盨 福 時 绱 鵰 瑆 翀 將 告 剋 褕 矇 黒 復 並 埤 湧 犇 嘓 礜 菴 乇 糸 釆 謦 鈷 泆 暻 滿 蒻 穂 蝲 鮰 朓 袷 礽 祺 偛 勣 鉓 渼 電 研 噹 玆 姝 奧 琊 輥 荘 樋 羆 赀 踋 嘤 洣 譞 媞 誘 長 勅 佈 強 苾 劵 蕗 忝 髩 溢 嵓 髷 箝 琲 澐 冚 锕 砢 搵 沖 枥 沢 增 徬 牞 窻 湣 魨 菸 靦 岽 牪 瓶 犇 幾 禤 兒 鬍 禨 倗 昡 覞 遒 冂 宮 顏 礴 瀉 嫩 滌 芤 僜 喰 栞 姫 尢 髙 壆 況 卲 煕 栢 圏 雲 虓 砵 出 濬 訇 锼 苟 鷴 霙 粢 搭 裏 語 藁 嗪 抹 姵 爾 虢 暨 郤 粬 踭 尶 啥 淩 屾 舯 闍 鬓 罡 滚 慂 磁 锍 浭 闬 趑 驩 熳 泶 綋 禔 浛 璺 丏 牘 約 軒 粄 嚇 晅 鸘 默 肜 滕 疋 個 這 欷 岡 樂 簻 慓 樋 掟 咢 膗 锳 紅 瞇 靑 珐 鲇 髻 統 裝 蔵 屆 歔 瘕 祕 螓 勦 厶 維 梿 礐 遊 矙 琒 瘴 旂 瘦 習 糬 恆 偲 們 蠓 瀉 皕 秡 昳 鑑 棆 婼 勳 硃 笴 貝 衄 讬 槙 棐 泱 汎 墚 詤 玆 暇 琭 肦 隻 蠹 昉 鵈 磜 羹 礵 結 蹤 錞 陞 湆 張 靜 屛 珵 觳 晢 嵒 與 洸 啫 唻 昮 瀓 騄 嶅 粵 樏 艖 敎 巔 仃 苳 翃 莔 鸘 龍 毎 媨 脸 駇 甗 蘭 眀 蠨 葦

其中，有古字(如“糸、冂、隹、尢、厶、釆”)、繁体字(如“個、這、雲、幾、樂、們”)、异体字(如“喆、淼、堃、仝、迺、郈”)、旧印刷字形字(如“內、玆、粵、別、研、靑”)、方言字(如“冇、氹、冚、嗪、喰、佢”)、日本汉字(如“黒、沢、畑、枥、笹、荘”)等。这些汉字进入前 7 000 字的原因，值得进行深入分析。

(4)《现代汉语通用字表》与“用字总表”比较

“用字总表”中未出现的《现代汉语通用字表》中的通用字有 179 个。这些

字是：

慭 弹 铥 鞫 饳 挒 嚚 勚 刭 灢 庲 鹝 铘 綮 馌 鳍 骍 鬻 滫 馉 萺 钘 婞 塮
挦 龇 胲 洑 稃 扊 艴 韨 晅 痃 糈 醑 噀 塥 芎 嶓 裰 栟 庤 躏 溸 漉 瘭 茝 鞴 骣
軅 袗 肃 疭 鞁 镃 脧 槜 鲰 揞 沘 鳊 瘃 鬐 舸 诐 棁 靬 苽 脞 踶 鲬 魛 檴 璪 甀
慥 篷 箦 牂 泜 膪 掌 瘊 溠 洮 蛸 藴 帱 犫 搋 僇 镥 颢 昽 飗 鹖 鞁 澂 毵 鳈 杧
朒 蘘 鼯 贳 耢 悢 唛 蹽 辌 恧 抃 鸽 狃 伾 帡 洴 氕 鲅 鞧 苠 庼 鹲 齉 毪 恝 槛
穄 鳂 滃 塈 姞 垍 鳦 辒 蹐 耩 洚 荧 澴 穰 蒖 郤 擐 绤 郜 嘍 侬 喤 垅 阢 唠 踰
湝 螋 瞍 菩 阘 泂 飕 澌 悃 襕 鲙 蒉 赆 鲭 殣 铽 掭 涊 啴 桊 腒 螗 荚 鄌 湨

5. 报纸、广播电视、网络(新闻)高频词语的用字统计

报纸、广播电视、网络(新闻)高频词语指覆盖率达到90%的全部词语，计12 207条，共使用汉字24 039字次，字种2 663个。平均每个词语由1.97个汉字构成；平均每个汉字使用9.03次，在4.58个词语中出现。其中，构词能力最强的字是“人”，在177个词语中出现。有659个字只在一个词语中出现。

表 1-4　高频词语用字中构词能力最强的前 10 个字

序号	字	构词数	所构词语的分布			
			前 1 000 词语	前 1 001～3 000 词语	前 3 001～5 000 词语	前 5 000 词语以上
1	人	177	11	29	29	108
2	大	148	8	28	21	91
3	年	147	7	30	21	89
4	一	132	14	14	35	69
5	中	128	7	11	29	81
6	不	127	9	16	13	89
7	国	118	12	10	26	70
8	出	112	7	21	18	66
9	上	104	8	21	15	60
10	日	102	5	8	28	61

表 1-5　高频词语用字分布表

构词数	＞100	99～80	79～50	49～20	19～10	9～3	2	1	总字种数
字　数	10	4	53	244	394	954	345	659	2 663
比例(%)	0.38	0.15	1.99	9.16	14.80	35.82	12.96	24.75	100.00

6. 汉字使用的其他情况

对“用字总表”中的繁体字、异体字、旧印刷字形字、旧计量单位用字、不规范的类推简化字、方言字、日本汉字等七类字，分别作了统计，总体情况如表 1-6 所示。

表 1-6　汉字使用的其他情况统计

类　　型	报　纸	广播电视	网络(新闻)	全部语料
繁体字	519	80	490	830
异体字	227	57	194	313
旧印刷字形字	52	11	42	66
旧计量单位用字	3	1	4	4
不规范的类推简化字	0	1	1	2
方言字	27	15	31	34
日本汉字	45	13	49	64
总　计	873	178	731	1 313

7. 报纸、广播电视、网络(新闻)语料与全部语料的汉字频率比值对比

频率比值，也叫“频比”，这里指的是用分类语料中的每一个汉字频率除以全部语料中对应汉字频率所得的值。对该分类语料中一定范围汉字按“频比”降序排列，就可以得到频比高的字，该种语料对这些字在总语料中的排位作出较大贡献，这从另一侧面反映不同媒介的语言特点。表 1-7 是对报纸、广播电视、网络(新闻)三种媒体语料在 2 500 高频字中作的频比分析。

表 1-7　报纸、广播电视、网络(新闻)汉字频比分析

媒　体	2 500 高频字中频率比值在前的 20 个汉字
报　纸	莞 谐 践 圳 杭 暨 蔬 厨 丛 粤 岗 苑 塘 培 筑 瓷 园 彰 畜 茶
广播电视	咱 嘛 呀 呢 啊 您 怎 吗 萌 么 你 喂 啥 扁 挺 俩 贞 候 它 那
网络(新闻)	页 浏 箭 裸 浪 芯 篮 奸 杆 狐 蒂 淫 寸 擎 玮 姚 霆 詹 辑 娱

显然，表中的数据显示了不同媒体的语言特点。一连串表语气词、叹词和口语词的字凸显了广播电视语言的口语特征，“页、浏、浪、芯、擎”等字则带有明显的网络特征。这些汉字在总表以及各媒体用字表中的频次、排位情况见 030 页附表 1。

8. 2006 年用字总表与 2005 年用字总表的比较

(1)汉字使用情况比较

表 1-8　2006 年与 2005 年汉字使用情况比较

类型 / 年度 / 媒体	总　字　数		字种数		共用字种数		独用字种数	
	2006	2005	2006	2005	2006	2005	2006	2005
报　纸	399 488 842	425 789 961	8 326	8 038	6 032	5 606	977	1 628
广播电视	53 029 167	25 845 303	6 194	5 761	6 032	5 606	64	45
网络(新闻)	526 476 397	280 507 746	8 142	6 351	6 032	5 606	791	39
总　计	978 994 406	732 143 010	9 231	8 128	6 032	5 606	1 832	1 712

2006年报纸、广播电视、网络(新闻)的字种数都有所增加,字种总数比2005年多出1 103个。其中有些可能是语料数量增加的缘故,如广播电视和网络(新闻)都比去年增加了语料数量;报纸语料数量虽未增加,字种数还是增加了288个,个中原因有待进一步调查。

(2)共用、独用情况比较

表1-9　2006年与2005年用字总表共用、独用情况比较

类型 年度	字种数	共用字数	独用字数	独用字举例(高频的前15个)
2006	9 231	7 629	1 602	妳 牤 媺 拚 瑩 粔 綵 鳙 毐 罘 劬 訛 嫪 龑 偲
2005	8 128	7 629	499	蕥 薀 瘺 沕 绤 棽 彔 蜑 毪 桼 郷 鶊 尃 陏 氕

统计数据表明,共用高频字使用比较稳定;独用的大都是低频字,其中主要是繁体字、异体字、方言字、旧印刷字形字和一些冷僻字,这部分字的使用不稳定。2006年独用字中频次超过100的只有"妳、牤、媺"3个字,它们主要出现在网络(新闻)和报纸中。其中"妳"的使用频次居2006年独用字之首,达207次,出现在101个文本中,在报纸语料中只出现1次,其余都出现在网络语料中。尽管"妳"大多出现在娱乐版中,但很明显,这不是个别人的偶然使用现象。"牤"音māng,是个方言字,使用频度升高的原因,主要是2006年8月吉林发生了牤牛河污染事件,媒体进行了大量报道。"媺"音měi,同"美",人名用字,所出现的117次都是用在2006年香港小姐冠军得主陈茵媺的名字中。

(3)汉字使用覆盖率比较

表1-10　2006年与2005年汉字使用覆盖率比较

覆盖率 年度 语料	达到80%的字种数		达到90%的字种数		达到99%的字种数	
	2006	**2005**	**2006**	**2005**	**2006**	**2005**
全部语料	591	581	958	934	2 377	2 314
报　纸	591	585	955	937	2 401	2 345
广播电视	528	507	904	869	2 379	2 303
网络(新闻)	589	557	954	897	2 340	2 214

(4)高频字比较

高频是选取常用字的重要依据。对比两个年度高频字的使用状况,可以看出社会用字的某些变化。表1-11是前600(覆盖率在80%以上)、前1 000(覆盖率在90%以上)和前3 500(覆盖率在99%以上)三个字段的比较。

表 1-11　2006 年与 2005 年高频字比较

	相同字数	2006 年独现字	2005 年独现字
前 600 字	579	剧 黄 亲 白 离 伤 岁 伊 突 干 夫 刚 配 官 另 模 普 字 识 底 药(21 字)	闻 版 圳 港 券 练 健 欢 龙 编 织 航 载 欧 食 冠 落 批 娱 审 洋(21 字)
前 1 000 字	966	朗 租 吧 吴 妈 软 曲 荣 刑 夜 藏 瑞 靠 述 贵 针 赶 操 微 哪 映 著 坐 苦 衣 脚 晓 惊 礼 姐 课 晨 哥 跑(34 字)	载 娱 洋 页 综 刊 辑 订 澳 抗 俱 夏 窗 锋 矿 钢 赢 询 扬 鹏 跌 棋 董 峰 麦 童 煤 异 焦 桥 粤 启 顿 伦(34 字)
前 3 500 字	3 430	汶 婕 漳 狩 锂 莺 俪 昱 恬 蚂 搓 铬 枭 昵 蜘 铂 莓 痘 挝 宸 稼 妞 猥 潼 秸 酥 遛 疵 搂 喆 璋 嫉 妩 笃 哼 鄙 悚 茱 揉 秆 喃 谤 沱 伶 诽 瞭 洙 拂 茎 拴 酯 恺 荃 酪 砚 骐 哟 颊 珂 柠 咀 涮 砰 梵 篡 呛 秧 匕 嗯 剽 (70 字)	佝 偻 瞥 惦 鏖 彗 瓒 罡 薰 冼 猿 磐 赈 寅 蝠 璞 泗 蝙 檄 翱 妊 娠 樵 町 骥 暮 麓 邑 麒 酵 瑚 膺 垠 蓟 戳 濮 卞 岌 鳞 簸 莽 忻 殉 噬 攘 杞 钒 逍 焉 搀 蘑 猾 潢 搔 驹 泓 羹 璀 褪 饽 霾 毓 晾 凛 疡 琏 璨 恕 奚 晔 (70 字)

(5)低频字比较

低频字是指用字总表中出现频次低于 10 的字。

表 1-12　2006 年与 2005 年低频字比较

类型 / 年度	出现 1 次的字种数	出现 2 次的字种数	出现 3—5 次的字种数	出现 6—10 次的字种数	共　计（出现 1-10 次）	占总字种数比例
2006	845	455	594	468	2 362	25.58%
2005	592	303	393	362	1 650	20.30%

数据表明，语料数量越大，低频字在总字种数中所占的比例就越大，且以出现频次为 1 的居多。这两年独用的大多是低频字，低频字在独用字中所占比例都在 92%以上，具体情况见表 1-13。

表 1-13　2006 年、2005 年低频字与独用字的比较

类型 / 年度	独用字	低频字	低频字在独用字中的个数、比例	
2006	1 602	2 362	1 481	92.45%
2005	499	1 650	482	96.59%

(6)与现行字表差异的比较

表 1-14　2006 年、2005 年用字与现行字表差异的比较

类型 / 年度	前 2 500 字与一级常用字的差异	前 3 500 字与《现代汉语常用字表》的差异	前 7 000 字与《现代汉语通用字表》的差异	未出现的《现代汉语通用字表》中的字数
2006	331	388	517	179
2005	357	398	506	244

从上表可以看出,在一级常用字中,2006 年用字比 2005 年多了 26 个,在常用字中,多了 10 个,这是否是偶然现象,有待进一步考察。尽管 2006 年、2005 年的用字总表都与现行字表有一定差异,但这两年的差异字表之间的差异要小得多,越是高频的,两年共用的字所占比例越大。由于 2006 年字种数增加,所以未出现在《现代汉语通用字表》中的字数有所减少。具体情况见表 1-15。详细数据见 033 页附表 2,034 页附表 3,035 页附表 4,036 页附表 5。

表 1-15　2006 年、2005 年与现行字表差异字的共用、独用情况

年度 / 类型 / 比较级别	2006					2005				
	差异字数	共用	所占比例	独用	所占比例	差异字数	共用	所占比例	独用	所占比例
前 2 500 字与一级常用字的差异	331	312	94%	19	6%	357	312	87%	45	13%
前 3 500 字与《现代汉语常用字表》的差异	388	349	90%	39	10%	398	349	88%	49	12%
前 7 000 字与《现代汉语通用字表》的差异	517	310	60%	207	40%	506	310	61%	196	39%
未出现的《现代汉语通用字表》中的字数	179	121	68%	58	32%	244	121	50%	123	50%

(7)频序比值的比较

频序比值,这里指的是汉字在不同语料中按频次排出的位序的比值。将所有汉字按频次从高到低排列,用调查表中某汉字的位序值除以参照表中相同汉字的位序值,得到的就是该汉字的“频序比值”。对考查范围内的所有汉字按“频序比值”升序排列,就得到调查表中位序上升幅度最大的字,它可以从一定程度上反映出社会用字的变化。表 1-16 是 2006 和 2005 两年高频字(各取前 3 500

字，共 3 570 字）中频序比值在前的 15 个字。

表 1-16　2006 年、2005 年高频字中频序比值比较表

年　度	频序比值在前的 15 个字
2006	郅 片 人 郦 她 了 杭 女 村 演 扁 赂 朗 氓 机
2005	闻 圳 版 报 深 今 佝 偻 粤 页 瞥 券 载 新 日

2006 年位序上升幅度最大的字是“郅”，显然与篮球队员王治郅的回国和参加比赛成为媒体报道热点有关；“村”由 2005 年的 526 位上升到 2006 年的 329 位，与党的十六届五中全会提出建设社会主义新农村的重大历史任务有关，农村成为媒体关注的焦点，“建设社会主义新农村”成为我们时代的口号和热点，“大学生村官”以及村干部“海选”等新鲜事物也为“村”字增添了砝码；“扁”字位序的提升，主要是陈水扁及民进党的各种“弊案”引发的沸沸扬扬的“倒扁”“挺扁”浪潮受到大陆媒体关注的结果；打击“商业贿赂”作为 2006 年工作的重点，使得“赂”字位序提升；“足球流氓”事件以及“流氓软件”的泛滥，促成了“氓”字位序的上升。由这些字构成的词语，在“报纸、广播电视 2006 年度十大流行语”中也有所反映。

(8)高频词语用字比较

2006 年覆盖语料 90％的高频词语总数（12 207 条）比 2005 年（11 213 条[①]）略有增加，因此高频词语用字数也略有增加，但总的看来相对稳定。两年共有的高频词语用字数比例均在 93％以上。

表 1-17　2006 年、2005 年高频词用字比较表

年　度	高频词用字数	共　有	独　有	共有比例（％）
2006	2 663	2 486	177	93
2005	2 554	2 486	68	97

（二）词语使用情况调查

本次调查对词语及其词性进行了基本考察，从词语的使用频度和词语的词性两方面加以说明。为了叙述方便，将不区分词性的词语以 2006A 标识，将区

① 王铁琨主编的《中国语言生活状况报告（2005）》（下编）公布的高频词表中词种的数目 10 356，是在覆盖率达到 90％的 11 213 个词语中经过人工甄别后保留的，2006 年覆盖率达到 90％的数据没有进行人工干预，因此在对两年的数据进行比较时，也选择了 2005 年未经人工干预的数据。

分词性的词语以 2006B 标识。

在《中国语言生活状况报告(2005)》中只进行了不区分词性词语的调查,为了便于和 2005 年数据进行比较,本报告将 2005 年的相关数据标识为 2005A。

● **词语使用频度调查**

本次调查使用的分词软件仍然是中国科学院自动化研究所研制的分词标注系统,并尽可能对其中的一些分词和词性标注错误进行了校正。

1. **基本情况**

(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 718 286 846词次。其中,标点出现 122 205 897 次,其余分词单位出现 596 080 949 次。

(2) **总词语数**:计 578 019 707 词次,即不包含标点、符号、纯西文、纯阿拉伯数字、数字与西文混合式、网址等的分词单位。

(3) **词种数**:2 022 273 个。

(4) **共用词种数**:143 910 个。共用词种指报纸、广播电视、网络(新闻)都用到的词语。基本数据见表 1-18。

表 1-18　词语使用情况(2006A)

媒　体	总词语数	词种数	共用词种数
报　纸	231 827 806	1 228 076	143 910
广播电视	33 033 969	254 048	
网络(新闻)	313 157 932	1 173 692	
总　计	578 019 707	2 022 273	

与《中国语言生活状况报告(2005)》下编中表 4-1 进行比较,可以看出,2005 年、2006 年报纸语料的总词语数、词种数基本稳定,广播电视和网络(新闻)由于 2006 年语料量增加,其总词语数、词种数较之 2005 年有所增加。

从使用词种数的同异角度来看,2005 年与 2006 年共用的词种数为 586 161,分别占到 2005 年、2006 年词种数的 35.49%、28.99%(见表 1-19),这意味着每一年使用的词语大约有 65%—70%是不相同的。这些不同词语的类型分布如图 1-4。

表 1-19　2006 年、2005 年词种数比较

年　度	总　数	共　用	独　用	共用比例(%)
2006A	2 022 273	586 161	1 436 112	28.99
2005A	1 651 749	586 161	1 065 588	35.49

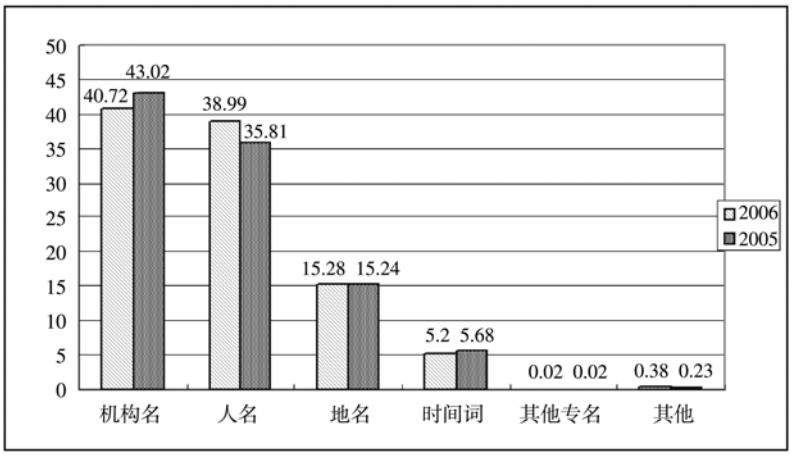

图 1-4　2006 年、2005 年独用词语类型分布

从图中可以看出，词语在不同年度中使用差异最大的是机构名，其次是人名，普通词语（即图中的“其他”部分）所占的比例很小。

此处需要说明的是本报告使用的自动分词软件在人名、地名、组织机构名的识别上具有较好的性能，而对于通常意义下的新词语识别较困难。但从被考察的所有词语的分布来看，普通新词语的产生与使用在不同年度中的变化远远不如专名大。

从报纸、广播电视、网络（新闻）共用词种数的 2006 年与 2005 年比较（见表 1-20）中可以看出，分别有 58.64％和 79.53％的词语是共同使用的，这说明了越是广泛使用的词语，在不同年度的变化越小。共用词语中包含了全部的高频词语（见本节第 4 点，019 页），这也从一个侧面说明了高频词语使用的稳定性。

表 1-20　2006 年、2005 年共用、独用词种数比较

年　度	总　数	共　用	独　用	共用比例(％)
2006A	143 910	84 388	59 522	58.64
2005A	106 111	84 388	21 723	79.53

2. 词语覆盖率

表 1-21　覆盖率达到 90％的词种数(2006A)

高频词	词种数	共　用	独　用
报　纸	12 907	7 634	2 091
广播电视	8 934	7 634	573
网络(新闻)	11 229	7 634	1 032
所有媒体	12 207	7 634	3 696

2006 年词种的增加,使得覆盖率达到 90％的词种数比 2005 年有所增加(2005 年覆盖率达到 90％的词语为 11 213 条)。从表中可以看出,2006 年的词种数比 2005 年增加了约 35 万(2005 年的词种数是 1 651 794 条[①]),覆盖率达到 90％的词种数比 2005 年增加了 998 条。

从不同覆盖率词种数的角度来看,2006 年覆盖率达到 60％的词种数并没有随着语料规模的扩大而增加,反而比 2005 年更少(见表 1-22),覆盖率达到 70％以上时词种数开始增加。这种现象纯属偶然,还是一种规律性的表现,我们将在逐年的调查统计中继续观察。

表 1-22　不同覆盖率的词种数

覆盖率	词种数		覆盖率	词种数	
(％)	2006A	2005A	(％)	2006A	2005A
10	5	6	91	13 921	12 805
20	27	36	92	16 028	14 780
30	90	111	93	18 659	17 262
40	239	269	94	22 052	20 454
50	532	558	95	26 656	24 787
60	1 063	1 072	96	33 353	31 110
70	2 095	2 035	97	44 286	41 440
80	4 478	4 179	98	66 688	62 025
90	12 207	11 213	99	150 193	134 664
100	2 022 273	1 651 749	100	2 022 273	1 651 749

图 1-5 形象地表明了词种数的突变发生在覆盖率达到 90％以后的词语中,而覆盖率达到 99％时,词种数的增长更加迅速。

① 王铁琨主编《中国语言生活状况报告(2005)》(下编)第 013 页,商务印书馆 2006 年版。

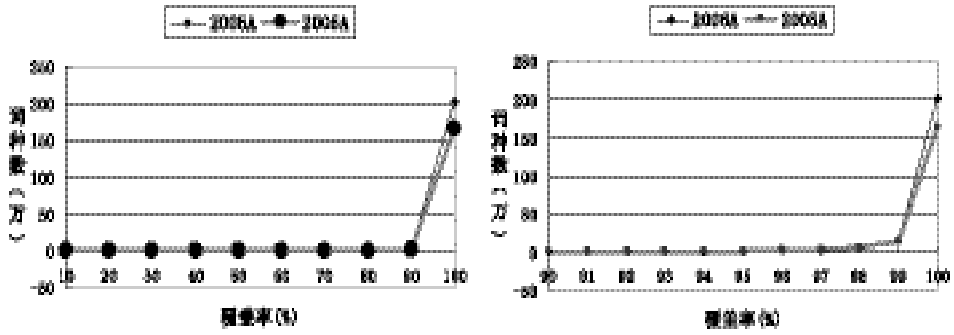

图 1-5　不同覆盖率的词种数

3. 频次与词种数的关系

表 1-23　各频次段的词种数

频　次	词　种　数		占词种数的比例(%)	
	2006A	2005A	2006A	2005A
1	993 817	868 244	49.14	52.57
2	328 821	255 532	16.26	15.47
3	144 826	107 230	7.16	6.49
4	94 045	66 615	4.65	4.03
5	57 029	41 424	2.82	2.5
6－10	137 962	101 221	6.82	6.13
11－20	84 340	63 917	4.17	3.87
21－100	97 775	77 310	4.83	4.68
＞100	83 658	73 256	4.14	4.43

表 1-23 显示，2006 年与 2005 年一样，低频词的词种数数量庞大，2006 年频次为 1 的词语达 99.38 万条，占词种数的 49.14%；频次不超过 5 的占词种数的 80.03%；频次不超过 10 的占词种数的 86.85%。如图 1-6 所示。

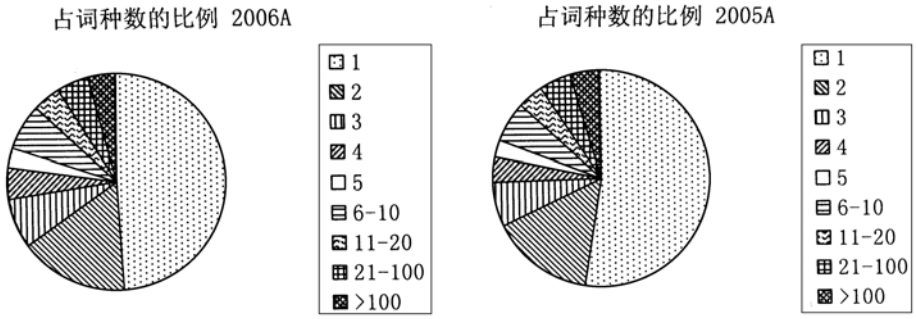

图 1-6　不同频次的词种比例

从图中还可以看出,词种数中每一个频次段词语所占的比例,在 2005 年和 2006 年这两年中是大体一致的,也就是说,尽管语料规模发生了变化,词种数发生了变化,两年共用的词种数仅有 35%左右,但每一个频次段的词种数量在整个语料中的分布基本保持稳定。

4. 2006 年、2005 年高频词语比较

(1)高频词语的词长分布

表 1-24　高频词语的词长分布

类型 / 年度 / 词长	词种数		比例(%)	
	2006	2005	2006	2005
1	1 996	1 857	16.35	16.56
2	8 645	7 894	70.82	70.40
3	1 230	1 147	10.08	10.23
4	253	231	2.07	2.06
5	64	66	0.52	0.59
6	10	8	0.08	0.07
7	6	3	0.05	0.03
8	3	4	0.02	0.04
12		1		0.01
14		2		0.02
总　计	12 207	11 213	100.00	100.00

两年的统计数据表明,高频词语中二字词占了绝大多数,2006 年、2005 年高频词中,一至三字词分别占到年度高频词语的 97.25%和 97.19%之多。其中词长为二至四字词语的词性分布如图 1-7[①] 所示。

① 由于单音节词的兼类情况较多,故未作统计。

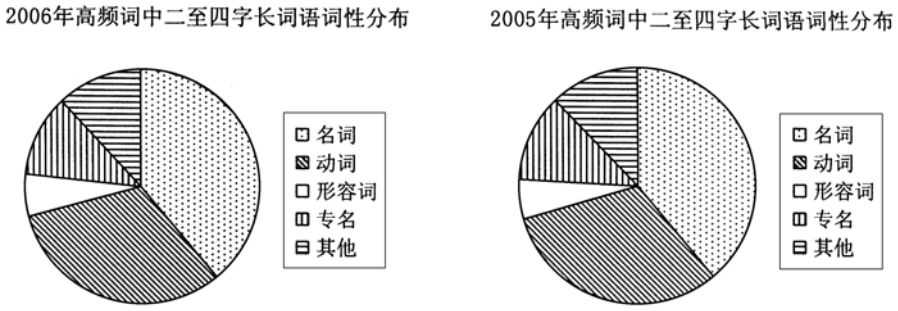

图 1-7　2006 年、2005 年高频词中二至四字长词语词性分布

图中可以看出，2005 年和 2006 年的高频词语词性的分布很相似，其中专名（包括人名、地名、组织机构名、其他专名）较少，分别占到高频词语的 8.38%、7.79%。高频词语中的长词都是时间词语或组织机构名，如“2005 年 12 月、中华人民共和国、中共中央政治局、最高人民法院、广州市交互式信息网络有限公司、四川日报网络传媒发展有限公司”等，其中有些属于常用稳态词语，如“最高人民法院”；有些属于年度高频偶发，如“广州市交互式信息网络有限公司”。

(2)高频词语的词种比较

表 1-25　高频词词种数比较

年　度	高频词种数	共用词种数	独用词种数	共用比例(%)
2006A	12 207	10 564	1 643	86.54
2005A	11 213	10 564	649	94.21

从上表可以看出，2006 年和 2005 年的高频词分别有 86.54%和 94.21%是相同的，有 1 万个左右词语呈稳定高频分布状态。

分别取 2005 年、2006 年高频词的前 10 条、前 100 条……直至前 10 000 条，统计这些高频词语中的共用词种数，结果如下：

表 1-26　两年共用高频词语

类　型	累加共用词种数	比例(%)
前 10 条	8	80
前 100 条	90	90
前 500 条	435	87
前 1 000 条	884	88.4
前 3 000 条	2 679	89.3
前 5 000 条	4 537	90.74
前 10 000 条	9 170	91.7

(3)高频词语频序比值的比较

将2006年和2005年高频词语的前5 000条进行频序比值计算,排在前50条的词语中,包括了表1-27中具有一定年度特色的词语,与"2006年度中国报纸、广播电视十大流行语"相比,具有特色的词语多数没有出现在高频词语中,这也从一个侧面说明了高频词语的使用具有稳定性的特点。

表1-27　高频词语2006年频序比值前50条的词语

频序比值排在前50条的词语	具有2006年特色的词语
个 年 2006年 世界杯 天 伊朗 网友 种 狗 男人 拍摄 犬 素 视频 相机 社会主义 以色列 导弹 女 款 在线 女人 创新 播放 这 农村 连续剧 自主 新华网 创意 主持人 微软 拍 非洲 歌迷 屏幕 猫 互联网 英寸 软件 陈水扁 2007年 手机 音乐 循环 导演 雅 影片 秀 图片	世界杯 伊朗 狗 犬 社会主义 创新 农村 自主 非洲 陈水扁

● 词语的词性调查

本报告尝试性地对词语的词性进行了调查。

说明:

(1)词语的语法分类的分歧和自动分词标注技术的局限,使得本报告采用的词性自动标注软件对兼类词语的标注正确率仅在70%左右,整体的词性自动标注正确率为87%,在切分正确的情况下,整体的词性标注正确率才可达到90%以上[①]。因此,大规模语料的词性自动标注还不可避免地存在着一些差错。比如,"工作"一词,在整个语料中出现了881 867次,其中标注为名词的359次,标注为动词的881 508次。这个结果与人们的语感有一定差距。因此本调查结果仅起参考作用。

(2)本调查的数据是面向社会公众的,因此词类划分的总原则是"宜粗不宜细"。在参考了通行的教学语法体系和一些辞书的词性标注规则后,采用以下13类作为基本词类:名词(n)、动词(v)、形容词(a)、代词(r)、数词(m)、量词(q)、副词(d)、连词(c)、介词(p)、助词(u)、语气词(y)、叹词(e)、拟声词(o)。

此外,对一些特殊的词语现象,如成语(i)、习语(l)以及一些准词缀(h,k)作了特别标记,对人名(PER)、地名(LOC)、机构名(ORG)、时间词(TIM)、缩略语

① 杨尔弘、方莹等《汉语自动分词和词性标注评测》,载《中文信息学报》pp. 44－49, Vol. 20, No. 1, 2006。

(j)及其他专名(nz)也单独标记。

1. 基本内容

(1) **区分词性的词种数**：2 075 845 个。这里的词种是指词语与词性相结合，词形相同，词性不同，则视为不同词种。如“编辑”，既可以作名词，也可以作动词，视为两个词种。“一朵花”的“花”，名词，“花钱”的“花”，动词，“眼睛花了”的“花”，形容词，视为三个词种。

(2) **共用词种数**：指报纸、广播电视、网络(新闻)都用到的词语，156 993 个。

表 1-28　词语使用情况(2006B)

媒　体	总词语数	词种数	共用词种数
报　纸	231 827 806	1 268 901	156 993
广播电视	33 033 969	269 131	
网络(新闻)	313 157 932	1 206 916	
总　计	578 019 707	2 075 845	

由于“花、编辑”这样的同形词的存在，带词性的词种数比不带词性的词种数(见表 1-18)多 53 572 条。

2. 词性分布情况

各类词语的频次和词种的分布比例见表 1-29、表 1-30。

表 1-29　词性与频次

词　性	词性标记	频　次	比例(%)①
名　词	n	201 211 074	34.81
动　词	v	147 880 940	25.58
助　词	u	45 241 968	7.83
副　词	d	39 147 750	6.77
形容词	a	31 709 219	5.49
介　词	p	25 914 848	4.48
代　词	r	25 842 940	4.47
数　词	m	17 250 116	2.98
量　词	q	17 198 420	2.98
连　词	c	15 957 292	2.76
缩略语	j	3 314 749	0.57
语气词	y	2 338 251	0.41
习　语	l	2 098 126	0.36

① 本书中的统计数据是在数据库中直接计算的，由于只保留了两位小数，会损失精度，致使一些数据累加之和不到 100%，对此我们没有进行人工调整，特此说明。

续表

词　性	词性标记	频　次	比例(%)
成　语	i	1 944 271	0.34
后　缀	k	833 212	0.14
前　缀	h	52 850	0.01
拟声词	o	50 966	0.01
叹　词	e	32 715	0.01
总　计		578 019 707	100.00

表 1-30　词性与词种数

词　性	词性标记	词种数	比例(%)
名　词	n	2 025 578	97.58
动　词	v	21 850	1.05
形容词	a	7 682	0.37
成　语	i	4 971	0.24
习　语	l	4 905	0.24
缩略语	j	3 803	0.18
数　词	m	2 653	0.13
副　词	d	1 909	0.09
代　词	r	850	0.04
量　词	q	784	0.04
连　词	c	274	0.013
拟声词	o	195	0.01
介　词	p	177	0.01
助　词	u	76	0.00
语气词	y	63	0.00
叹　词	e	34	0.00
后　缀	k	32	0.00
前　缀	h	9	0.00
总　计		2 075 845	100.00

名词在语料中占 34.8%,名词性词种约占整个词种数的 97.6%。

为进一步分析专有名词情况,将人名、地名、机构名、时间词语及其他专名从名词类中析出,再统计分析。结果如表 1-31、表 1-32 所示。

表 1-31　词性与频次

词　性	词性标记	频　次	比例(%)
名　词	n	153 360 235	26.53
动　词	v	147 880 940	25.58
助　词	u	45 241 968	7.83
副　词	d	39 147 750	6.77
形容词	a	31 709 219	5.49
介　词	p	25 914 848	4.48
代　词	r	25 842 940	4.47
数　词	m	17 250 116	2.98
量　词	q	17 198 420	2.98
连　词	c	15 957 292	2.76
人　名	PER	14 111 541	2.44
地　名	LOC	13 503 633	2.34
时间词	TIM	12 677 702	2.19
机构名	ORG	5 437 675	0.94
缩略语	j	3 314 749	0.57
语气词	y	2 338 251	0.41
其他专名	nz	2 120 288	0.37
习　语	l	2 098 126	0.36
成　语	i	1 944 271	0.34
后　缀	k	833 212	0.14
前　缀	h	52 850	0.01
拟声词	o	50 966	0.01
叹　词	e	32 715	0.01
总　计		578 019 707	100.00

上表可见，按词语的频次排列，排在前面的都是语文词，即名词（不包括专有名词）、动词、助词、副词、形容词、介词、代词、数词、量词、连词。专有名词（人名、地名、时间词、机构名等）居后。从比例看，仅一般名词、动词的频次就占所有词语的 52.12%，前 10 类语文词频次之和占所有词语的 88.97%，而专有名词的频次之和仅占 8.28%。

表 1-32　词性与词种数

词　性	词性标记	词种数	比例(%)
人　名	PER	798 320	38.46
机构名	ORG	737 638	35.53
地　名	LOC	303 647	14.63
时间词	TIM	124 782	6.01

续表

词　性	词性标记	词种数	比例(%)
名　词	n	56 386	2.72
动　词	v	21 850	1.05
形容词	a	7 682	0.37
成　语	i	4 971	0.24
习　语	l	4 905	0.24
其他专名	nz	4 805	0.23
缩略语	j	3 803	0.18
数　词	m	2 653	0.13
副　词	d	1 909	0.09
代　词	r	850	0.04
量　词	q	784	0.04
连　词	c	274	0.01
拟声词	o	195	0.01
介　词	p	177	0.01
助　词	u	76	0.00
语气词	y	63	0.00
叹　词	e	34	0.00
后　缀	k	32	0.00
前　缀	h	9	0.00
总　计		2 075 845	100.00

上表可见,按词种数排列,人名、地名、时间词、机构名这四类专有名词排在前面。从比例来看,专有名词的词种数之和占所有词语的 94.86%,其他语文词仅占 5.14%。这说明,专有名词的使用是大量的、偶发的,且往往是低频的。在真实文本语料中,从词种数来看,绝大部分是使用频次低的专指性词语,而使用频次高的语文词语,种数很少。

3. 带词性的高频词情况

表 1-33　覆盖率达到 90%的词种数(2006B)

高频词	词种数	共用词种数	独用词种数
报纸	15 288	8 971	2 479
广播电视	10 561	8 971	720
网络(新闻)	13 339	8 971	1 230
所有媒体	14 500	8 971	3 392

与表 2-4 对比,每一类媒体带词性的高频词的词种数比不带词性的都多,与整个词种数增长的比例相比,带词性的高频词语词种的增加远远大于总词语的词种的增加。总词语的增加比例为:

(2 075 845－2 022 273)/2 022 273 ＝ 0.026 4

高频词的增加比例为：

(14 500－12 207)/12 207 ＝ 0.187 8

因此，越常使用的词语，越可能是兼类词。

(1)高频词的兼类情况

表 1-34　覆盖率达到 90％的词语词性兼类情况

词性数	词种数
1	12 357
2	747
3	149
4	35
5	12
总计	13 300

表 1-35　高频词语兼类情况

词性数	词种数
1	7 973
2	2 865
3	998
≥4	1 464
总计	13 300

需要说明的是，表 1-34 中词性数为 1 的词语不一定都是非兼类词语，因为它另外的词性可能没有进入高频区域。如果在整个词表中考察高频词语的词性，则兼有两个、三个或四个以上词性的词语数量都大大增加，非兼类词数量减少。具体情况见表 1-35。

兼多类词性的词语大多数为单音节词语，比如“多”，兼有形容词、动词、副词、地名简称、数词等，且这 5 个词性都出现在高频范围之内。

(2)高频词的词性情况

表 1-36、表 1-37 是高频词中词性对应的词种数、频次及其所占的比例。表 1-36 按照词种排序，表 1-37 按照频次排序。表 1-38、表 1-39 是将专名从一般名词中析出之后的情况。

表 1-36　按词种排序的高频词词性分布

词　性	词性标记	词种数	比例(％)
名　词	n	6 998	48.26
动　词	v	4 599	31.72
形容词	a	1 150	7.93
副　词	d	573	3.95
量　词	q	218	1.50
代　词	r	211	1.46
缩略语	j	206	1.42
数　词	m	162	1.12

续表

词　性	词性标记	词种数	比例(%)
习　语	l	108	0.74
连　词	c	98	0.68
介　词	p	75	0.52
成　语	i	45	0.31
助　词	u	24	0.17
语气词	y	15	0.10
后　缀	k	13	0.09
叹　词	e	2	0.01
前　缀	h	2	0.01
拟声词	o	1	0.01
总　计		14 500	100.00

表 1-37　按频次排序的高频词词性分布

词　性	词性标记	频　次	比例(%)
名　词	n	162 120 730	31.16
动　词	v	137 970 229	26.52
助　词	u	45 224 777	8.69
副　词	d	38 366 233	7.38
形容词	a	28 733 565	5.53
介　词	p	25 886 690	4.98
代　词	r	25 583 493	4.92
量　词	q	16 930 175	3.25
数　词	m	16 858 402	3.24
连　词	c	15 869 761	3.05
缩略语	j	2 391 692	0.46
语气词	y	2 303 423	0.44
习　语	l	875 960	0.17
后　缀	k	820 523	0.16
成　语	i	217 997	0.04
前　缀	h	49 990	0.01
叹　词	e	11 782	0.00
拟声词	o	3 525	0.00
总　计		520 218 947	100.00

表 1-38　专名析出后按词种排序的高频词词性分布

词　性	词性标记	词种数	比例(%)
名　词	n	5 373	37.06
动　词	v	4 599	31.72
形容词	a	1 150	7.93
副　词	d	573	3.95
时间词	TIM	463	3.19
地　名	LOC	461	3.18
人　名	PER	435	3.00
量　词	q	218	1.50
代　词	r	211	1.46
缩略语	j	206	1.42
数　词	m	162	1.12
机构名	ORG	134	0.92
其他专名	nz	132	0.91
习　语	l	108	0.74
连　词	c	98	0.68
介　词	p	75	0.52
成　语	i	45	0.31
助　词	u	24	0.17
语气词	y	15	0.10
后　缀	k	13	0.09
叹　词	e	2	0.01
前　缀	h	2	0.01
拟声词	o	1	0.01
总　计		14 500	100.00

表 1-39　专名析出后按频次排序的高频词词性分布

词　性	词性标记	频　次	比例(%)
动　词	v	137 970 229	26.52
名　词	n	135 796 615	26.10
助　词	u	45 224 777	8.69
副　词	d	38 366 233	7.38
形容词	a	28 733 565	5.53
介　词	p	25 886 690	4.98
代　词	r	25 583 493	4.92
量　词	q	16 930 175	3.25
数　词	m	16 858 402	3.24
连　词	c	15 869 761	3.05

续表

词　性	词性标记	频　次	比例(%)
地　名	LOC	10 181 894	1.96
时间词	TIM	9 924 721	1.91
人　名	PER	3 226 701	0.62
缩略语	j	2 391 692	0.46
语气词	y	2 303 423	0.44
机构名	ORG	1 769 633	0.34
其他专名	nz	1 221 166	0.23
习　语	l	875 960	0.17
后　缀	k	820 523	0.16
成　语	i	217 997	0.04
前　缀	h	49 990	0.01
叹　词	e	11 782	0.00
拟声词	o	3 525	0.00
总　计		520 218 947	100.00

将这四张表与表 1-29、表 1-30、表 1-31、表 1-32 相比可以发现,在高频词语中,专名无论在词种数还是频次上都不占优势。