中文博客用字用语专项调查

博客创造了一种新的以个人为中心的传播方式，成为继电子邮件、BBS 论坛和即时聊天之后的第四种网络交流方式。为了了解博客语言使用状况，国家语言资源监测与研究中心网络媒体语言分中心建立了中文博客语料库，对其用字用语情况进行了调查统计。

一　语料说明

本次调查的对象是在中国大陆注册的中文博客网站。根据网络访问量、网站规模等因素，选择了 6 个知名的中文博客网站：blog. sina. com. cn；blog. sohu. com；blogcn. com；bokee. com；blog. hexun. com；blogbus. com。从中下载了 2006 年的部分网页，共计 401 768 个文本，135 910 409 字符次。

博客语料库中记录了博客作者和博客文章的相关信息，记录格式如下：

表 4-1　博客作者格式

作者名	网 址	XML 链接	标 题

表 4-2　博客文章格式

标 题	网 址	正 文	发表时间

二　调查内容

调查项目包括博客中汉字、词语、符号的使用情况，并将这些情况与《报纸、广播电视、网络（新闻）用字用语调查》（见本书第 001～036 页）进行了初步比较。

三　调查结果

(一)基本情况

本次调查统计了23 520个博客作者创作的文本,文本总数为401 768个,文本的平均长度为1 250个字符,文本长度的分布情况见表4-3。

表4-3　文本长度分布

文本长度(字符)	文本数	百分比(%)	累加百分比(%)
0—63	118 095	29.39	29.39
64—127	63 540	15.82	45.21
128—511	89 306	22.23	67.44
512—1 023	33 036	8.22	75.66
1 024—2 047	25 558	6.36	82.02
2 048—3 071	17 391	4.33	86.35
3 072—4 095	13 660	3.40	89.75
≥4 096	41 182	10.25	100.00

(二)汉字使用情况

1. 总汉字数:计108 709 287字次,约占总字符数的80%。

2. 汉字种数:11 923个。

3. 汉字覆盖率见表4-4,字种数和覆盖率的变化曲线见图4-1。

表4-4　汉字覆盖率分布

覆盖率(%)	50	80	90	99	100
字种数	128	586	1 041	2 868	11 923

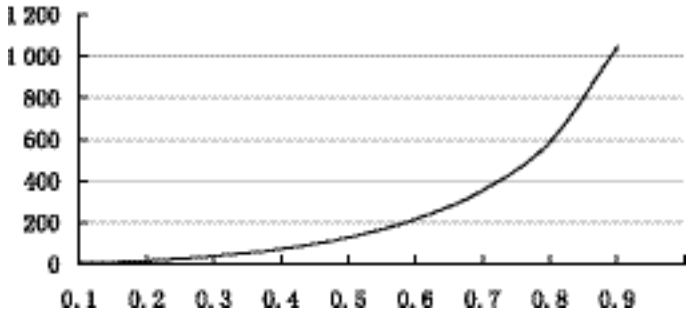

图4-1　汉字覆盖率分布

4. 在 11 923 个汉字中，繁体字出现了 1 616 个，占汉字种数的 13.55%；异体字出现 504 个，占汉字种数的 4.23%。这些字使用频率低，它们在不同字种范围的分布见表 4-5。

表 4-5　繁体字、异体字分布

字种数	繁体字种数	异体字种数
前 1 000	0	0
前 2 000	5	0
前 3 000	36	1
前 5 000	340	30
前 7 000	819	134
11 923	1 616	504

(三)词语使用情况

与 2006 年其他几个调查项目一样，我们分别统计了区别词性和不区别词性条件下词语的使用情况。本报告中如不特别指明，一般是在区别词性的条件下统计的数据，把词形相同、词性不同的词条看作不同的词条。

1. 基本数据

(1)总词语数：71 698 903 词次。

(2)词种数：425 929 条。其类别分布情况见表 4-6 和图 4-2。在本报告的表格、图表中，“其他”栏主要指名词、动词、形容词、副词等语文词语。

表 4-6　词语类别分布

词语类别	词种数	所占比例(%)
人　名	168 675	39.60
地　名	45 958	10.79
机构名	48 024	11.28
时间词语	23 458	5.51
数字词语	39 779	9.34
其　他	100 035	23.48

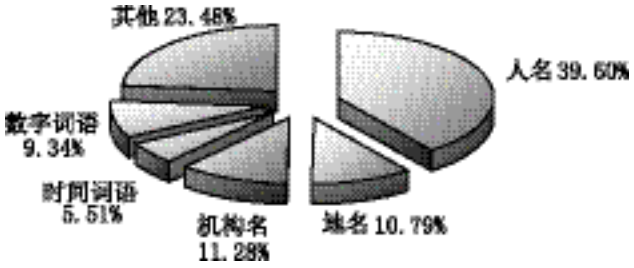

图 4-2　词语类别分布

使用频次为 1 的低频词语词种数为 217 903 条，其类别分布情况见表 4-7。

表 4-7　低频词语类别分布

词语类别	词种数	所占比例(%)
人　名	104 598	48.00
地　名	30 666	14.07
机构名	36 582	16.79
时间词语	13 706	6.29
数字词语	24 147	11.08
其　他	8 204	3.76

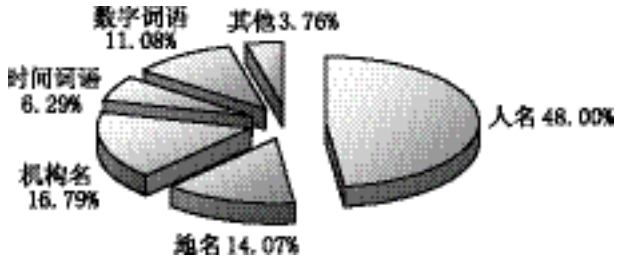

图 4-3　低频词类别分布

2. 词语覆盖率

覆盖率百分之八十以上的各种覆盖率下的词种数见表 4-8。

表 4-8　覆盖率百分之八十以上的词种数

覆盖率(%)	80	90	91	92	93	94	95	96	97	98	99	100
词种数	4 563	13 243	15 072	17 270	19 973	23 351	27 689	33 486	41 788	55 335	87 180	425 929

表 4-9 显示了百分之八十和百分之九十覆盖率下的词类分布状况。

表 4-9　高频词类别分布

覆盖率(%)	词种数	人名	地名	机构名	时间	数字	其他
80	4 563	11	46	2	101	147	4 256
90	13 243	148	146	23	319	331	12 276

3. 频次与词种数的关系

表 4-10　各频次的词种数

频 次	词种数	所占比例(%)
1	217 903	51.16
2	48 270	11.33
3	21 059	4.94
4	13 448	3.16
5	9 094	2.14
6－10	23 958	5.62
11－20	20 238	4.75
21－100	37 562	8.82
＞100	34 397	8.08

4. 词种的文本分布情况

表 4-11　各文本数段的词种数

文本数	词种数	所占比例(%)
1	243 285	57.12
2	42 936	10.08
3	18 807	4.42
4	11 280	2.65
5	7 817	1.84
6－10	20 557	4.83
11－20	17 816	4.18
21－100	33 836	7.94
＞100	29 595	6.94

(四)符号使用情况

符号总数:指除汉字以外的符号的总次数,计 27 201 122 字符次，占总字符次的 20.01%。

符号种数:928 个。

符号使用的基本情况见表 4-12。

表 4-12　非汉字符号基本情况

类　别	出现次数	出现种数	占语料库比例(%)
标点	14 925 586	126	10.98
数字	2 320 277	106	1.71
字母	5 186 548	454	3.82
其他	4 768 711	242	3.5

(五)博客用字用语与报纸、广播电视、网络(新闻)用字用语比较

与 2006 年《报纸、广播电视、网络(新闻)用字用语调查》相比,博客的用字用语情况具有以下特点:

1. 汉字的使用更具有多样性,主要表现在:

(1) 博客的字符种数、汉字种数明显高于报纸、广播电视、网络(新闻),博客的汉字种数高达 11 923 个,特别是繁体字的字种数和使用频次明显增加。

(2)博客覆盖率达到 80%的字种数略少于报纸、广播电视、网络(新闻),但覆盖率达到 90%以上的字种数却多于后者。具体数据见表 4-13。

表 4-13　不同覆盖率的汉字数对比

覆盖率(%)	汉　字　数	
	报纸、广播电视、网络(新闻)	博客
80	591	586
90	958	1 041
99	2 377	2 868
100	9 231	11 923

2. 词语的使用数量有所减少,主要表现在:

(1)博客词语的覆盖率情况与报纸、广播电视、网络(新闻)有些不同,在不区分词性的条件下,前 69 657 个词覆盖率达到 99%,比后者达到相同覆盖率所用的词数少了 80 536 个。具体数据见表 4-14。

表 4-14　不同覆盖率的词种数对比

覆盖率(%)	词　语　数	
	报纸、广播电视、网络(新闻)(2006A)	博客
10	5	2
20	27	13
30	90	42
40	239	109
50	532	283
60	1 063	668
70	2 095	1 498
80	4 478	3 462
90	12 207	9 894
91	13 921	11 282
92	16 028	12 961
93	18 659	15 026
94	22 052	17 637
95	26 656	21 034
96	33 353	25 632
97	44 286	32 254
98	66 688	43 151
99	150 193	69 657
100	2 022 273	1 173 692

(2) 在区分词性的条件下,博客专有名词词种数所占的比例与报纸、广播电视、网络(新闻)相比,也有些不同,其中机构名的比例只有 11.28%,大大低于 2006 年报告的 35.53%,具体数据见表 4-15。

表 4-15　专有名词词种数比例对比

词性	所占比例(%)	
	报纸、广播电视、网络(新闻)(2006B)	博客
人名	38.46	39.60
机构名	35.53	11.28
地名	14.63	10.79

(3)高频词语的使用各有特点。博客语言中,“我、我们、自己、你、她、他”之类的人称代词的使用频率比报纸、广播电视、网络(新闻)高。另外,“去、想、说、知道、喜欢、觉得、吃、让”等常用动词的使用频率也明显高于后者,而“中国、记者、新闻、市场、企业、公司、国际、本报”等和新闻报道相关的词语使用频率明显低于后者。