首都平面媒体用字用语状况调查[1]

曾小兵 杨尔弘 张普

北京语言大学应用语言学研究所 北京 100083

E-mail:xiaobingzeng@126.com

摘 要: 本文选取了北京地区的10种主流报纸建成首都地区平面媒体动态流通语料库，对其中的用字、用语情况进行多角度多层次的考查与数据分析，涉及字词的频次、频率、使用率、覆盖率、构词能力、频比等多项指标，并尝试将首都地区平面媒体动态流通语料库与国家语言资源监测语料库平面媒体库进行一些比较，用计算语言学的研究方法在大规模真实语料中考查首都地区语言生活的“实态”，探求其内在特征及外部规律。

关键词: 首都地区 平面媒体语料库 语言生活

The Investigation on the Chinese Characters and Words' Usage of Print Media Language Corpus in Capital Area

Zengxiaobing,Yang erhong,Zhangpu

Applied Linguistics institute in BLCU Beijing 100083

E-mail:xiaobingzeng@126.com

Abstract: This paper selected 10 mainstream newspapers in Beijing and established the Dynamic Circulation Corpus of the Print Media, trying to make a survey with data analysis on the Chinese characters and words' usage from the multi-angle and hierarchical ways, which involves many indexes as the frequency, the utilization rate, the coverage rate, the formation capacity of words, the frequency ratio, etc. We attempt to compare the Dynamic Circulation Corpus of Beijing with the Print Media Language Corpus of National Language Resources, then show the actual language life using computational linguistics methods which based on the large-scale real corpus , and explore its internal features and external rules.

Keywords: The capital Area, Corpus of the Print Media , Language life

语言生活是社会生活的一个不可或缺的组成部分，它一直倍受语言学专家、学者的关注。“从语言生活的历史进程看，人类语言生活的发展节奏越来越快”，这已经成为许多学者的共识（王均，2000[1] 李宇明，2004[2] 王铁琨，2007[3] 周有光，2008[4]）。为了动态地对语言生活的“实态”进行调研与总结，国家语言监测与研究中心课题组已经连续发布了三个年度的《中国语言生活状况报告》，本文尝试在此思想的指导下，以首都地区为切入点，依据大规模真实语料（平面媒体）对其用字用语情况进行调查分析，从中探究首都地区的语言生活特色。

1. 语料及调查说明

(一) 调查使用的语料

经媒体的流通度及可获得性等因素考虑，选取了北京地区的10种主流报纸[2]作为调查语料。前期的调研发现，由于社会的全面进步与信息的迅猛发展，仅供某地区的居民阅读的报纸是极少的，故而我们将只在北京地区发行或绝大部分在北京地区发行并供北京居民阅读的报纸媒体作为调查对象，从候选的24份报纸中提取上述10份。调查语料共计439,654个文本文件，901,167,219

1本文得到北京市教育委员会共建项目“汉语国际推广背景下的首都留学生教育研究”资助.

2 这10种报纸是（按音序排列）：《北京科技报》、《北京青年报》、《北京青年周刊》、《北京日报》、《北京晚报》、《法制晚报》、《京华时报》、《青年周末》、《新京报》、《中国教育报》

字符次（包括标点、符号及西文字母、数字等出现的次数），其中汉字出现213,609,505字次。

（二） 调查内容

我们主要是从汉字和词语的使用情况入手，考查并反映出首都地区的语言生活。既从首都地区语料库内部进行考量，也从首都地区语料库（北京地区10家报纸）和国家语言资源监测语料库的平面媒体库（全国范围内15家报纸）之间进行比较来考查。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、出现文本数及使用率的含义及计算方法可参考《中国语言生活状况报告（2007）》（下编）的“语言资源监测与研究相关术语”。[5]

2. 汉字使用情况

（一） 基本情况

汉字总数：指全部字符中去除汉字部件、乱码、无法显示的字符以及仅在“火星文”中出现的书写符号后，计213,609,505字次；字种数：共7,345个，所有字种形成调查的用字总表。

（二） 汉字的覆盖率

汉字的覆盖率即指定的汉字占所有汉字总量的比例。汉字覆盖率是反映汉字整体分布与总体使用情况的重要指标。其统计结果见下表1-1。

表1-1 北京语料库中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
报纸	604	976	2 422	7 345

自上个世纪初开始，“艾思杜（J.Estoup）、贡东（E.Condon）、齐普夫（G K.Zipf）、朱斯（M.Joos）、曼德尔布洛特（B.Mandelbrot）等人先后研究了词的出现频率与词的序号之间的相互关系。”[6]而覆盖率则是将频率与序号有机结合后的考查，即按调查对象的频次降序排列后计算其累加频率。从表1-1的汉字分布结果来看，较少的字种数覆盖了语料的大部分，而99%—100%的范围内则包含了4,923个字种，占总字种的67.03%。这也正符合了zipf（1949）等人所说的省力原则（the principle of least effort）[7]，用较少的汉字来表达或描述较多的事物。这也是语言的发展演变规律之一。

（三） 北京地区语料库汉字使用的规范性

将北京地区语料库用字总表前2500字与《现代汉语常用字表》[3]一级常用字（2500字）比较，用字总表中有330字是一级常用字中所没有的。它们多表示人名或北京地区的地名，如：淀、颐、蔡、曹、彬、斌、邓、迪、蒂等。将用字总表前3500字与《现代汉语常用字表》（3500字）比较，用字总表中有382字是《现代汉语常用字表》中所没有的，其中有些表示突发事件，如：汶、胺等。将用字总表前7000字与《现代汉语通用字表》（7000字）比较，用字总表中有638字是《现代汉语通用字表》中所没有的，这些字多具偶然性，并没有太多规律。用字总表中未出现的《现代汉语通用字表》中的字有482个。此外，用字总表中的繁体字（99个）、异体字（90个）、不规范的类推简化字（5个）、旧计量单位用字（2个）、日本汉字（23个）、方言字（16个）等，出现在用字总表的2 851位以后，即覆盖率达到99.48%的汉字范围之后。具体情况见表1-2。北京地区的汉字使用情况总体来讲，与现行的汉字规范呈现较强的一致性，但也出现了上述一些不规范、错误的使用，这是多方面的原因造成的，这使我们进一步认识到，语言的

3国家语言文字工作委员会、国家教育委员会1988年联合发布。

规划及语言政策的调整，应该在语言发展的客观规律及实际状态的基础之上进行。

表 1-2 用字总表与规范用字表的比较

范围	一级常用字之外的字数	范围	常用字表之外的字数	范围	通用字表之外汉字数
前 500	1	前 1000	1	前 3000	2
501-1500	48	1001-1500	3	3001-5000	32
1501-2500	281	1051-2500	77	5001-7000	604
–	–	2051-3500	301	–	–

（四） 北京地区语料库与国家语言资源监测语料库平面媒体库用字的比较

国家语言资源监测语料库平面媒体库（以下简称平面媒体语料库）是国家语言监测与研究中心发布年度语言生活状况报告的基础语料库。我们将北京地区语料库（以下简称北京语料库）与 2008 年的平面媒体语料库进行比较，以期发现北京地区语言生活的一些特色。

1. 共用、独用情况

表 1-3 北京语料库与平面媒体语料库汉字独用、共用情况

类型	数量	频次降序排在前 10 位的
共用	7 058	的 一 在 是 人 了 中 有 国 不
平面媒体语料库独用	1 098	珧 鐽 咭 笹 韡 溇 实 蜞 冪 蘭
北京语料库独用	2 87	芏 詨 蕡 錩 鎛 垹 鰭 轀 鰛 楻

对于北京地区独现的前 18 个汉字，我们调查了其具体的语言环境，其中“楻、獳、轀”主要是引用性的文字，在文中分别来源于《天工开物》、《左传》、《史记·秦始皇本纪》；作为方言使用的有“腡、啕、褐”分别来自儿歌《手腡歌》、拟声词“噢啕”和方言词“汗褐子”（指衬衣）；“鰭、鰛”是用作稀有动物的名称：“小鰛鲸”、“黄金鰭鱼”；“芏、垹、僪、棃、込、渥、鎛、詨、蕡、錩”主要用于人名。我们还发现，北京地区独用的汉字数量较少、使用的频次较低、出现的文本数少，使用的偶然性大，没有必然的规律。但从中可以看出，独用字以人名的使用居多。

2. 频率比值

频率比值（简称频比）能够在一定程度上反映不同媒介、不同领域语言的特点。[8] 表 1-3 分别列出了北京语料库与 2008 年平面媒体语料库中前 2500 字范围内的频比值排在前 20 位的汉字。

表 1-4 北京语料库与平面媒体语料库汉字频比分析

媒体	前 2500 字中频率比值在前的 20 个汉字
平面媒体语料库	圳 深 广 舰 州 粤 莞 省 业 氰 宝 江 杭 骡 市 禹 荔 钇 五 旗
北京语料库	京 北 薷 昨 教 癜 校 版 编 淀 演 剧 赛 闻 访 请 颛 癣 羌 讯

从上表可以看出两者的差异，平面媒体语料库涉及的领域和地域广泛，“圳、深、州、粤、莞、宝、江、杭、禹”等字充分反映了其与北京不同的地域性特点，北京语料库中“京、北、淀”则充分反映了北京的特色，“演、剧、赛”与北京地区多样的文化活动相关联，“闻、访、请”等与北京地区的政治性、外交性活动频繁有关，而“校、版、编”等则是与北京地区语料库中部分报纸的性质相关，新闻版面信息多，短小精悍、数量增加使得“校对、编辑”等字眼增多。

3. 高频字

由于高频字有强稳定性，其在北京语料库和平面媒体语料库之间的异同，能够更多更有效地说明两者的差别与共性。在此选择前 600（覆盖率在 80%以上）和前 1000（覆盖率在 90%以上）两个范围段进行比较，统计结果见表 1-4。

表 1-5 北京语料库与平面媒体语料库高频字比较

比较范围	相同字数	北京地区独现字
前 600 字	567	版 编 闻 访 彩 馆 试 伤 音 卡 像 列 它 觉 存 纪 练 言 愿 致 航 离 绍 铁 曾 普 故 夫 班 语 停 飞 击 （33）
前 1000 字	964	赵 晨 孙 亡 童 哈 迷 毒 诺 射 延 述 阵 禁 夺 杯 穿 杰 微 迪 梦 探 遭 呼 爆 圣 烟 染 萨 炬 塔 授 订 杀 饭 异（36）

可以看出，两者共用的高频字占了绝大多数，即高频字在地域、领域的分布具有稳定性。从北京地区独现的字也可以看出北京地区的一些特点，如："版 编 闻 访"等，这些特点在前 600 字范围内有体现，当范围再扩大时，其特色体现不明显。

4. 低频字字种与比例

低频字是指用字总表中出现频次低于 10 的字。在北京语料库中进行统计，具体情况见表 1-5。

表 1-6 北京地区语料库与平面媒体语料库低频字字种数

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	合计	占总字种数比例（%）
北京语料	558	241	393	305	1 497	20.38
平面媒体	725	321	434	330	1 810	22.19

上表数据表明，低频字在两个语料库中的总数及占有的比例相对一致，但在具体的出现次数的字种数有较大的差别，其中原因有待于进一步探究。

3. 词语使用情况

（一） 基本情况

总词语数：共计 125,458,033 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。词种数： 718,955 个。

（二） 频次与词种数的关系

表 1-7 不同频次范围的词种数

频次	词种数	所占比例（%）	累计（%）	频次	词种数	所占比例（%）	累计（%）
1	347 446	48.33	48.33	6-10	47 253	6.57	83.31
2	110 016	15.30	63.63	11-20	32 214	4.48	87.79
3	45 220	6.29	69.92	21-100	45 181	6.28	94.07
4	30 726	4.27	74.19	>100	42 617	5.93	100.00
5	18 282	2.54	76.73				

表 1-6 列出了不同频次范围的词种数情况。从表中可以看出，频次不超过 5 的词种数占

76.73%，频次不超过 10 的词种数占 83.31%，即低频词的词种数数量巨大。

（三） 词语的覆盖率

表 1-7 列出了 10%到 90%，91%到 100%各段覆盖率的词种情况。从表中可以看出，词种数的明显上升是在覆盖率为 90%之后的词语中，这与汉字的分布情况极为类似。

表 1-8 不同覆盖率的词种数

覆盖率(%)	词种数	比例（%）	覆盖率(%)	词种数	比例(%)
10	6	0	90	12806	1.78
20	34	0	92	16658	2.32
30	105	0.01	94	22631	3.15
40	267	0.04	96	33418	4.65
50	583	0.08	98	62484	8.69
60	1153	0.16	100	718955	100
70	2275	0.32			
80	4806	0.67			

（四） 高频词语

1. 基本情况

词语覆盖率达到 90%的所有词语称为高频词语。高频词语的词种数为 12,806 个。

2. 高频词语用字统计

在 12,806 个高频词语中，共使用汉字 25,190 字次，2,764 个字种，占全部字种数的 37.63%。平均每个词由 1.967 个汉字构成。每个汉字平均使用 9.114 次。这些数据说明高频词语的用字情况是相对稳定的。其构词的情况见表 1-8，其中有 678 个字只在一个词中出现。

表 1-9 高频词语用字分布

构词数	＞100	99-80	79-50	49-20	19-10	9-3	2	1	总字种数
字数	10	7	51	267	416	955	380	678	2764
比例（%）	0.36	0.25	1.85	9.66	15.05	34.55	13.75	24.53	100.00

表 1-9 列出了在高频词语中构词最多的 10 个汉字及其构词数量。从表中可以看出，这些字所构成的词语数量上差别不大，高频词的用字情况也稳定。

表 1-10 高频词语用字中构词能力最强的前 10 个字

汉字	人	大	年	一	中	国	日	不	出	上
构词数量	186	157	156	135	123	121	116	114	109	103

3. 高频词语的词长分布

表 1-11 高频词语的词长分布

词长	1 字	2 字	3 字	4 字	5 字	6 字	7 字	8 字	9 字	总计
词种数	2031	9125	1296	251	81	10	8	3	1	12806
比例（%）	15.86	71.26	10.12	1.96	0.63	0.08	0.06	0.02	0.01	100

从表 1-10 中可以看出，词长为 4 字以上的词语累计占到高频词语的 99.20%，而字长较长的高频词语多为时间词（如 9 字词是：2008 年 1 月 1 日）和专有名词（如地名、组织机构名等）。

4. 高频词语的独用情况

表 1-12 北京语料库与平面媒体语料库高频独用词排序前 20 个

年度	高频独用词排序前 20 个
北京语料库	京华 责编 美编 短信 采写 丰台区 昌平 北京站 李章洙 李冬美 丰台 顺义 圆明园 辅路 中院 适房 八达岭 西城区 布托 李安
平面媒体语料库	粤 招租 品类 我市 中专 服务区 西门子 博世 省委 滨海 互动式 我省 新区 2008 年 让利 大餐 诚聘 珠三角 羊城 特区

5. 高频词语中的成语使用情况

表 1-13 2008 年高频词语中成语的分布

	词种		词次	
	数量	在高频词语中的比例（%）	数量	占高频词语词次比例（%）
成语	25	0.20	27763	0.02

6. 北京语料库与平面媒体语料库高频词语中的频序比值情况

表 1-14 北京语料库与平面媒体语料库高频词语中的频序比值情况表

年度	高频词语中的频序比值排序前 20 个
北京语料库	访问 本版 精彩 校 京 北京电视台 编 通州 朝阳区 海淀 教师 海淀区 京城 北京市 嫦娥 网 来京 大兴 请 北京地区
平面媒体语料	五星 大专 旗舰 深圳 作废 遗失 宝 五星级 厂房 佛山 家电 升级 广州 深圳市 东莞 广州市 成都 年龄 光明 工程师

以上对高频词的调查一方面可以反映高频词在语言中的稳定性。另一方面，我们也可以看到高频词等语言现象在反映社会生活中的作用。“语言的功能不是简单地等同于语言的使用，而是组织语言体系的基本原则。”[9] 因此，从语言的使用中去研究语言的功能及体系是可行的，但并不是最全面的，我们应该将语言的使用、功能及社会结构综合起来看语言生活的特点与规律。

4. 结论

北京作为我国的首都，是历史悠久的古都，传统政治文化的中心，有着深厚的文化积淀和丰富的语言文化资源。因此，北京地区的语言文化现象早就引起了人们的关注并已有细致的研究。如胡明扬等在 1987-1989 年就组成“北京话研究”课题组对北京话“做一些基础性的调查，做了一点进一步开展北京话研究的准备工作。”[10] 白公、金汕等从文学角度研究北京语言，“没有任何一种地方方言与地域特色文学联系得如此紧密而独树风格的”[11] “北京口语语料库”建成后，高海洋、彭宗平等人都据此开展了大量的研究。程祥徽、范德博、周庆生、杨晋毅、戴庆厦等专家学者也逐

渐关注城市或少数民族的语言生活并进行了多方面的调研。我们基于大规模真实语料库对首都的语言生活进行了一些基础性的调查，谈不上太多的创新，而且限于书面语语料的局限性，一些口语性的语言特征反映得不够明显。但利用计算机技术处理大规模真实语料而得到客观真实的数据，进而反映语言生活，这就如控制论中的“白箱”与“黑箱”：我们用“看得见、摸得着”的媒体这个“白箱”，去和语感这个“黑箱”进行相似性类比,将白箱的量化数据传递到黑箱。[12]这项工作本身就是一种充满挑战性的尝试和极具应用性的研究。

另外我们也应该看到，语言生活在稳定中有变化而且对于社会有十分重要的作用。索绪尔曾指出语言符号的“不变性”和“可变性”。由于时间因素和社会力量的效力，语言符号回到理性的“连续性原则”上，“但连续性必然隐含着变化，隐含着关系的不同程度的转移。”[13]这种可变性近年来在社会语言学领域受到空前的重视，Weinreich，Labovt等（1968）就指出语言是“异质有序”的。这种“异质”本身是一种有规律的语言现象，而这种异质的“有序”则表明了语言的变化反映了其社会地位和角色的差异。[14]Halliday也认为，语言不但具有反映社会现实的功能（概念功能），而且还能积极地作用于社会（人际功能），能够创造社会现实，使人类既能用语言创造社会意义，也可以用语言参加各种社会实践。[15]

因此，如何在城市化进程中关注迅速发展变化的语言生活，利用最高效可信的方法、手段进行城市及言语社区的语言生活调查，是值得社会语言学家及计算语言学家进一步思考的问题。

参 考 文 献：

[1] 王均 网络时代的语言生活和语言教学 语文建设 2000 年第 10 期

[2] 李宇明 促进语言生活健康发展 语言文字应用 2004 年 11 月

[3] 王铁琨 计算机统计数据与年度语言生活状况报告 长江学术 2007 第 1 期

[4] 周有光 语言生活的历史进程 徐州师范大学学报(哲学社会科学版) 2008 年 3 月 43-46

[5] 国家语言资源监测与研究中心编《中国语言生活状况报告（2007）》下编，P516-523 页，商务印书馆 2008 年）

[6] 冯志伟，齐普夫定律的来龙去脉 情报科学 1983 年第 2 期

[7] zipf Human Behavior and the Principle of Least Effort:An Introduction to Human Ecology.Cambrige, MA: Addison Wesley Press

[8] 国家语言资源监测与研究中心编《中国语言生活状况报告（2007）》下编，P9 页，商务印书馆 2008 年）

[9] M.A.K.Halliday 著 语言与社会 北京大学出版社 2007 年 11 月

[10] 胡明扬等 著《北京话研究》北京燕山出版社 1992 年 3 月

[11] 白公 金汕著《京味儿——透视北京人的语言》 中国妇女出版社 1993 年

[12] 张普 关于控制论与动态语言知识更新的思考 语言文字应用 2002 年第 1 期 p71-76

[13] 费尔迪南·德·索绪尔 著 高名凯 译 《普通语言学教程》 商务印书馆 2005 年 P116

[14] Weinreich.U,W.Labov&M.Herzog.1968.Empirical foundations for a theory of language change. In W. Lehman and Y. Malkiel (eds.)Directions for a historical linguistics. Austin: University of Texas Press.)

[15] 胡壮麟 朱永生 张德禄 李战子 《系统功能语言概论》 北京大学出版社 2008 年 P345