共和国建立以来党报用字用词的计量调查*：
——以《贵州日报》和《人民日报》为例

饶高琦[1,2]

北京语言大学1 语言科学院、2 信息科学学院 北京 100083

E-mail: raogaoqi@blcu.edu.cn

摘　要：对超过五十年的中文历时语料进行的计量研究尚属少见。本文对由《贵州日报》和《人民日报》构成的历时语料库（65 年跨度）进行了用字用词的基础信息考察。发现就语料总体而言用字用词规律性强，长尾现象严重。而用字用词的次数、种数随时间变化基本稳定，变化趋势呈现不对称U 型：文革前略高于文革期间，并少于文革后。这一变化趋势也与字词熵值的变化相符。时段独用字、词则可以较好的反映当时各时段的语言生活。

关键词：用字 用词 语言监测 社会语言学 历时语料

Computational Investigation in Words and Characters usage of Official Newspapers since Foundation of PRC: *Guizhou Daily* and *People Daily* as Examples

RAO Gaoqi[1,2]

1 Language Science Institute, 2 College of Information Science, Beijing Language and Culture University, Beijing 100083

E-mail: raogaoqi@blcu.edu.cn

Abstract: Computational research focused on diachronic corpus over half century range is rare. We investigated the word and character usage in the diachronic corpus consisted by Guizhou Daily and People Daily. “long tail” phenomena is obvious in word and character usage in whole corpus. Amounts of types and tokens of words and characters are roughly stable, while their trends in 65 years shows an asymmetric U: the amount of types and tokens in pre culture revolution is slightly higher than that in culture revolution, and lower than that of post culture revolution. Similar trends was also found by the calculation of information entropy. Words and characters appeared only in one specific period performed well in presenting temporal language life.

Keywords: word usage, character usage, social linguistic, diachronic corpus, language monitoring

1　前言

随着语言信息处理和语料库技术的日益发达，社会语言监测工作取得了很大的进步。从2004年起，教育部语信司发布《中国语言生活状况报告》已历十年，并通过对字、词、语的量化统计发布《语言生活状况数据篇》。我们认为不限于对当前语言生活进行研究，对过去语言生活的“考古”同样也是一种语言监测。但是超过十年的大时间跨度的中文语言历时定量考察，除刘长征[1]在三十五年深圳日报上的研究和 LIVAC 的两岸三地语料库系列研究[2]（十六年跨度），在社会语言学、媒体语言学研究中还很少见，而超过五十年跨度的计量研究则更加少见。

共和国成立以来到 20 世纪末的时间段，由于网络媒体还处于萌芽中，有声媒体的公开数据较少，中央、地方党报的语料则成为观察大时间跨度语言生活的一个重要窗口。本文使用了《贵州日报》和《人民日报》的语料，总跨度 65 年[3]。通过对语料基础信息：用字、用词情况进行分析，可以观察到我国语言生活的变迁的一个大致脉络。就语言使用的丰富程度而言，总体呈现了文革前略高于文革期间而大大弱于文革后的现象。本文第三、四章将从用字和用词的角度进行详细分析。

*本文受到国家自然科学基金项目（61300081，61170162）；国家社科重大基金项目（12&ZD173）；国家语委科研基金项目（YB125-42）；北京语言大学研究生创新基金（14YCX074）的资助。

2 语料情况

本文采用了北京语言大学信息科学学院收集的自共和国建立以来《贵州日报》和《人民日报》两份党报共 8 亿字的语料。从 1950 年（1949 年 11 月贵州省解放，贵州日报从 12 月开始发行，因此语料规模过小忽略）到 2014 年的语料规模分布如下图所示。年均语料规模 1237.5 万字。该语料 2013 年以前部分已经加工成为全文检索索引，在线开放使用[1]。

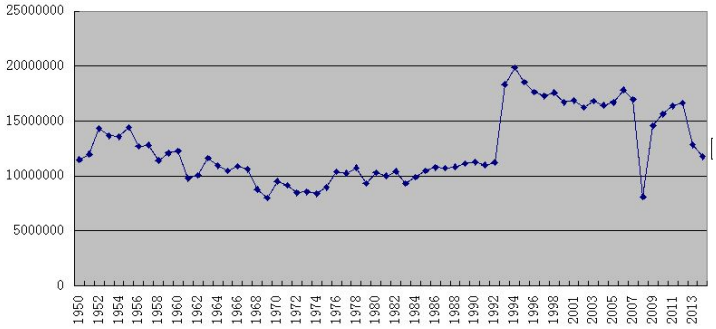

图 1 各年份语料规模（字数）

其中《贵州日报》的覆盖范围是 1950 年至 2007 年，《人民日报》的覆盖范围是 2009 年至 2014 年。两者均为党报，语体相同，内容和形式上具有很好的同质性。两份报纸的交替也没有改变语料规模变化的基本态势。后文诸统计涉及的拐点和变化点亦都不与语料交界点重合。

我们将 65 年的语料以文化大革命时期（1966-1976）为界，分为文革前（1950-1965 年）、文革时期（1966-1976 年）和文革后三段（1977-2014 年）。各时间段语料规模和年平均语料规模如图 2 所示。

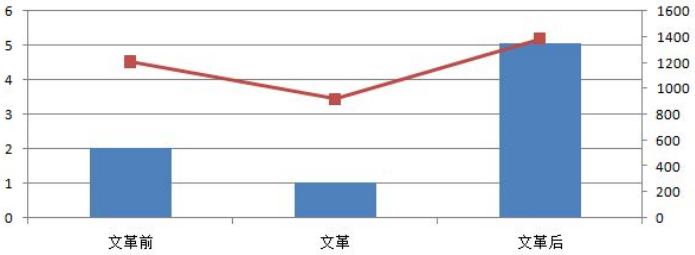

图 2 各个时间段语料量（柱为总字数/亿，点线为年均字数/万）

从图 1 和图 2 中可以看出，语料总体呈现 U 字形，即文革时期报刊篇幅较小，语料规模（约一亿字，年均九百余万字）略小于文革前（总计两亿字，年均一千两百万），大大小于文革后（约五亿字，年均将近一千四百万字）。

3 用字情况

3.1 字种与字次

表 1 显示了 65 年语料中各时期语料所使用的字种与字次情况（包括全半角阿拉伯数字和标点符号）。全部八亿字次的语料中，所使用的字种不足八千字（7824 字）。字种数与字次数的分布情况如图 3 所示。

时间段		文革前	文革	文革后	全部时段
该时段总字种数		6622	5989	7604	7824
该时段年均字种数		5029	4674	5517	5234
该时段总字次数		194865104	101759117	507753697	804377918
该时段年均字次数		12107356	9250829	13723073	12375045
各年间共用字	字种数	3336			
	比例	50.4	55.7	43.9	42.6
	字次数	194147290	101349686	503120880	798617856

[1] 在线语料库地址为 http://nlp.blcu.edu.cn/others/historical%20computing/

	比例	99.6	99.6	99.1	99.3
时段间共用字	字种数	5791			
	比例	90.2	96.7	76.1	74.0
	字次数 比例	194859510 99.997	101758167 99.999	507034363 99.858	803652040 99.910
时段独用字	字种数 比例	153 2.3	50 0.83	1028 13.5	1231 15.7
	字次数 比例	752 0.00039	94 0.000092	623058 0.12	623904 0.078

表 1 各时段用字情况

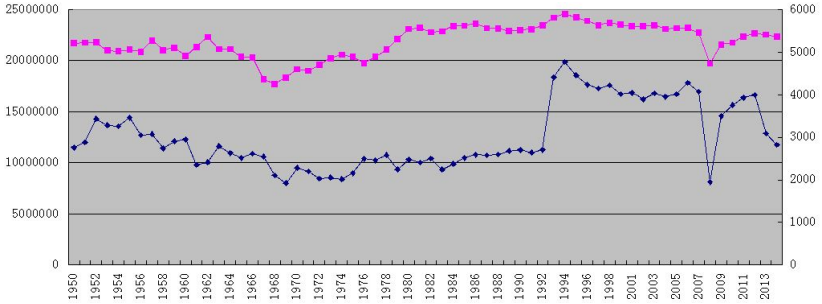

图 3 字种数与次数分布（字次数为左轴蓝色点线，字种数为红色点线右轴，单位字）

通过语料考察，可以推测共和国成立以来的用字水平整体稳定，但字次数和种数的变化趋势呈不对称的 U 型，即文革前的字种数和次数略高于文革期间而明显低于文革后。

整体稳定表现为字种数的标准差为 387 字，不足字种数最低年份（1968 年 4251 字）的百分之十。同时各年间共用字，占到总字量的 42.6%。各时段的共用字则高达 74%。这也从另一个侧面说明了用字量的总体稳定性。而这些共用字在实际使用中又占有绝对优势：各年共用字的字次数占总体 99.3%，各时段共用字的字次数达到了 99.9%

不对称 U 型则表现为，平均而言三个时期的字种数差异为五百上下，即文革中较之文革前少用了约五百字，比文革后少用了约一千字。这从一个侧面体现了文革时期语言使用的情况。

同样的情况也表现在字熵的统计结果上。熵是度量系统复杂程度的统计指标[4]，其公式如下所示：c_i 为字表中的第 i 个字，n 为字种总数，p 为由最大似然估计得到的字概率。而熵值增加从一个侧面表明了语言使用丰富程度的增加。各年度语料字熵见图 2。

$$Entropy = -\sum_{i=1,c_i \in C}^{i=n} p(c_i)\log p(c_i)$$

公式 1 信息熵（字熵）

图 4 各年度字词熵（柱为字熵右轴，词熵为红曲线，去停用词后为蓝曲线，均为左轴，单位 bit）

图 4 中字熵也表现出和字次数、种数相类似的变化趋势。通过熵，不对称 U 型的变化趋势更加明显。文革后用字量增加，汉字使用更加丰富、多样导致熵值大幅提高。

3. 2 汉字使用分布规律性强

从表 2 中可以估算语料的平均字频次为十万次，但是实际上汉字使用频率差异很大。覆盖率是体现汉字使用分布的一个指标，揭示了字种数与字次数之间的关系。通过对语料的考察可以发

现大时间跨度的用字分布呈现长尾的现象，即绝大多数频率由少数高频字构成。

覆盖率	达到 80%		达到 90%		达到 99%		达到 100%
	字种数	比例	字种数	比例	字种数	比例	字种数
文革前	457	6.90%	773	11.67%	2083	31.46%	6622
文革期间	415	6.93%	728	12.16%	1994	33.29%	5989
文革后	500	6.58%	833	10.95%	2275	29.92%	7604
全部语料	495	6.32%	830	10.61%	2261	28.90%	7824

表 2 各时段用字覆盖度

传统观点认为，高频的 600 左右汉字可以覆盖语料的 80%[5]，但是从表 2 显示这个估计仍然偏高。500 高频字足以覆盖语料的 80%，5563 个低频字只覆盖了语料的 1%。还注意到，文革期间的用字情况，达到 80%、90%和 99%所需的字量比例都较文革前和文革后时期高。这说明文革时期用字的不均衡性较其他时期略弱。

3.3 各时段独用字

各时段的独用字使用情况里，文革后时期的种数和次数均压倒前面两个时期。其中频次最高的是十个全角阿拉伯数字与百分号和小数点。之所以如此是因为在前两个时期，这些符号的功能均由汉字来完成，即%均写作“百分之”，数字和日期均用汉字等。后面的独用字则有科技术语用字如“熵”、“幂”、“镧”等，科技用符号如希腊字母、数学符号，民俗宗教词语中的用字如“貔”、“貅”、“祐”，外语符号（拉丁语族、日语符号）等。这些都反映了我国在文革后科教兴国、学习世界、保护传统文化方面的所做的努力。

与之构成强烈反差的就是文革前时期的独用字大部分是注音符号如“ㄙ”、“ㄛ”等，这也反映了《汉语拼音方案》公布以前我国的语言生活状况。

同时注意到相较于高频字，时代独用子更好的反映了那个时期的语言生活情况。因为三个时期的高频字均为“的、了、一、在、国、人”等没有时间敏感性的基本常用汉字。

4 用词情况

4.1 词种与词次

		文革前	文革	文革后	全部时段
该时段总词种数 该时段年均词种数 该时段总词次数 该时段年均词次数		1577160 92774 116980339 6916938	688587 62598 59271870 5465100	4558359 123199 300746626 8128287	6824106 104986 476998835 7338444
各年间共用词	词种数	11788			
	比例	0.75	1.71	0.26	0.17
	词次数 比例	106189258 90.8	53466569 90.2	256612651 85.3	416268478 87.3
时段间共用词	词种数	108313			
	比例	6.7	15.7	2.4	1.6
	词次数 比例	115296677 98.6	58621954 98.9	287622622 95.6	461541253 96.8
时段独用词	词种数 比例	306154 19.4	75095 10.9	889716 19.5	1270965 18.6
	词次数 比例	691859 0.59	186535 0.31	7586814 2.5	8465208 1.8

表 3 各时段用词情况

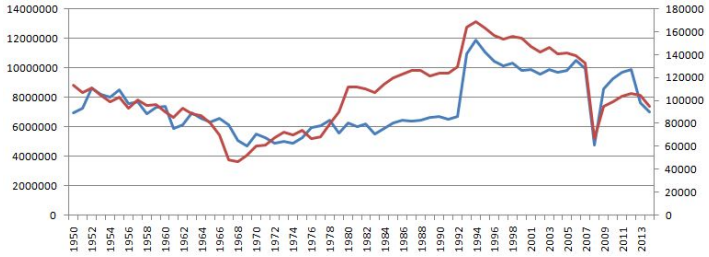

图 5 词种数与词次数分布（词次数为蓝曲线左轴，单位词）

我们在语料中统计了各时段的用词情况。词语调查是在对语料进行自动分词的情况下进行的，使用的分词工具是北京语言大学通用分词系统 GPWS[6]。各时段的用词情况如如表 3 所示。各年度的词次数与词种数分布如图 5 所示。在 65 年的语料中总词量约为六百八十万，词次数约为四亿八千万次，对比八亿字的总字次数，则平均词长为 1.67 字。

和用字情况相类似，65 年来的用词情况也呈现整体稳定，变化呈不对称 U 型的趋势。整体稳定表现为各年间共用一万余词（仅占词表的 0.17%）覆盖了全部语料的近 90%。而占词表 1.6%的十万词就足以覆盖全部语料的近 97%，在各时段也都在 95%以上。而 U 型变化则表现为文革前时期的词种和词数均略高于文革时期而明显落后于文革后时期。其变化幅度较用字情况明显许多。用词情况也可以使用熵来进行度量。图 3 显示词熵的两种分布。其中停用词为经验获取的，对时间不敏感的常用词（如：活动、关系、食品等等，频率通常很高），去停用词后的熵值的绝对值发生变化但变化趋势依旧。

4. 2 词语使用分布规律性强

从表 4 中可以估算语料的每词平均频次为 70 次，但实际情况是少部分高频词覆盖了大部分语料。平均而言词种数的 1.7%就覆盖了总语料的 99%，而想要看懂全部语料的 80%只需要掌握不足三千词就可以了。这对汉语教学具有一定的参考意义。

覆盖率	达到 80%		达到 90%		达到 99%		达到 100%
	词种数	比例	词种数	比例	词种数	比例	词种数
文革前	2035	0.13%	5844	0.37%	71045	4.5%	1577160
文革期间	1581	0.23%	4483	0.65%	42024	6.1%	688587
文革后	3179	0.07%	9250	0.20%	113902	2.5%	4558359
全部语料	2980	0.04%	8832	0.13%	113298	1.7%	6824106

表 4 各时段用词覆盖度

4. 3 词长

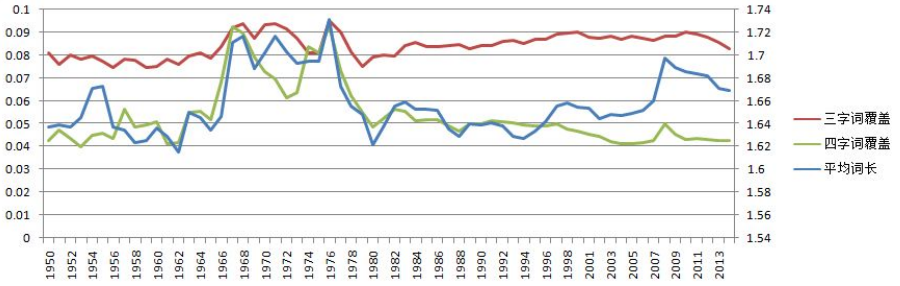

图 6 平均词长和多音节词对语料的覆盖（平均词长为左坐标轴）

图 6 统计了语料中的平均词长和双音节与三音节词覆盖语料的情况。值得注意的是文革时期平均词长出现了明显的增长，另外两个明显增长则出现在 1954 年和 2008 年。二字词对语料的覆盖相对稳定，而三字词和四字词对语料的覆盖与平均词长的变化呈现了很高的相关性。词长与二字词覆盖率的相关系数为-0.075，而与三字词、四字词覆盖率的相关系数为 0.649 和 0.655。平均词长的增长主要由二字以上的词语频繁使用贡献而来。

共和国建立早期政治运动（1954-1955）和文革时期新概念新提法数量猛增，这是导致二字以上词覆盖率增加，平均词长增长的重要原因。如 1954 年的“社会主义”、“人民公社”、“互助社”等，又如文革时期的“无产阶级文化大革命”、“革委会”、“走资派”等。与之相类似的 2008 年“汶川地震”和“奥运会”等词的高频使用也拉高了平均词长。

4.4 时段独用词

时段独用词中，文革后时期占了较大比例。表 4 显示了文革后独用词最高频的十个词。文革后时期语料延伸到 2014 年，故所包含的最后一届执政十年的政府是胡温政府。再此之前若干届政府的首脑均在文革中和文革前就展露政坛（江泽民和胡耀邦在文革和文革前时期就见诸报端），因而没有成为独用词。“改革开放”、“经贸”、“媒体”等词则反映了文革后时期巨大的社会变革。前十名中唯一的动词是“弘扬”（首现于 1987 年），在文革和文革前时期类似语境中通常使用“发扬”。并且注意到随着“弘扬”使用词频的提升，“发扬”的频次逐渐下降，如图 7。

和用字情况一样，较之高频词（去除标点符号后），独用词能够更好的反映时代特征。如表 5 所示。“的”“和”“了”等由于频次极高而排名靠前，但是由于其虚词和助词属性，对语义刻画的帮助较小，自然也没有时代特征。

胡锦涛	改革开放	习近平	温家宝	乡镇企业	经贸	媒体	残疾人	武警	弘扬
胡传魁	朗诺-施里玛达(集团)	韩小强	赵勇刚	钱守维	高志扬	多中心论	唐成模	马洪亮	赵震山

表 4 文革后时期和文革时期独用词频率前十位（第一行为文革后时期）

独用词	胡锦涛	改革开放	习近平	温家宝	乡镇企业	经贸	媒体	残疾人	武警	弘扬
高频词	的	和	了	在	一	是	为	发展	个	中

表 5 文革后时期独用词和高频词（高频前十）

文革中高频独用词大多为专有名词。一个有趣的现象是“重要配角现象”，即在文革时期的语料中某一重要事件的“配角”会成为独用词。如“胡传魁”是样板戏《沙家浜》中的男配角（主角如阿庆嫂和刁德一在文革后依然和沙家浜这部作品紧密联系在一起），“韩小强”、“赵勇刚”都来自样板戏《海港》，并且也是配角。而《海港》本身也是几大样板戏中知名度较低的一部。“朗诺-施里玛达.”和“朗诺-施里玛达集团”则是柬埔寨 318 政变的发动者（罢黜西哈努克国王，使其流亡中国），后为波尔布特的红色高棉政权消灭，也可以被认为是柬埔寨政局中的“配角”。而“多中心论”则是文革时期与“毛泽东同志为中心”相悖的，被批判的思想，可以视为那一时期思想斗争中的“配角”。

其实“重要配角现象”可以获得一种较为简单的解释，即在重大事件中起重要作用的“主角”还会在后一个时期的语料中频繁出现，如四人帮及其成员的名字。而偶然使用的词语又不足以进入独用词的高频段，因此高频独用词会呈现出一种“重要配角现象”。这一现象在文革前时期的独用词中也十分明显。从这一角度上来看，由于“主角”并不出现在的高频独用词中，因而“重要配角现象”也制约了高频独用词对时代语言生活的刻画。

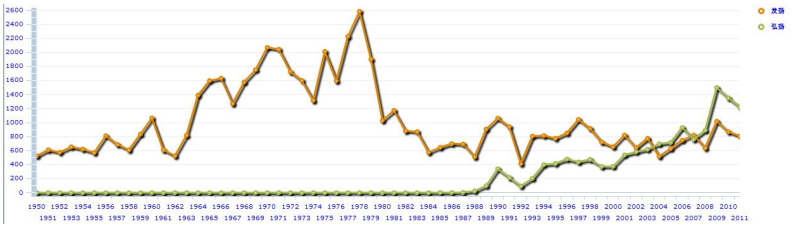

图 7 “发扬”与“弘扬”在语料中的分布（词次数）

5 结语与展望

我们通过对 65 年语料中用字用词频次和熵的统计和分析，发现共和国建立以来党报用字用词呈现总体稳定，变化趋势呈不对称 U 型的现象。用字用词均呈现很强规律性，极少数高频字词

覆盖绝大多数语料，频率分布比通常的估计更加不均衡。

独用字词在反映语音生活方面有较好表现。独用字、词在反映语言生活上的作用比高频字、词更强的这一现象，与自然语言处理中 TFIDF[2]特征可以进行文本分类相符合[7]。因为对不同时代语料的描写本身就可以视作对不同类型文本的分类任务。而独用字词本身就可以看做蜕化了的 TFIDF 特征。如果适当放宽标准，以频率曲线走势来寻找能够反映时代特征的词汇，则可以克服“主要配角现象”对刻画时代特征的制约。

针对大时间跨度语料的研究还有很多方面，本文只是就用字用词进行了基础调研。用语层面、命名实体的分布以及句法和构式方面都研究还亟待开展。

参 考 文 献

[1] 刘长征. 基于动态流通语料库的新词语监测研究[M] 北京：世界图书出版社.2011.

[2] 邹嘉彦，邝蔼儿，陆斌，蔡永富. 汉语共时语料库与追踪语料库[J]. 中文信息学报. 2011 年 11 月 25 卷第六期 P38-45.

[3] 饶高琦，荀恩东，谢佳莉，黄志娥. 现代汉语词汇历时检索系统：基于长时间跨度语料库的词汇历时信息研究与应用[C]. 第十四届国际汉语词汇语义学研讨会 CLSW-2013：郑州

[4] 克劳德·艾尔伍德·香农: 《通信的数学理论》 (A mathematical theory of communication) [J]. 贝尔系统技术月刊l,27 卷,379-423

[5] 教育部语言文字信息管理司：《中国语言生活状况报告(2013)》[M] 北京：商务印书馆.2013 P195.

[6] 宋柔，罗智勇.现代汉语通用分词系统(GPWS v3.5）http://democlip.blcu.edu.cn:8081/gpws/

[7] 宗成庆. 统计自然语言处理[M] 北京：清华大学出版社 P423.

[2] TF-IDF（Term Frequency Inverse Document Frequency），自然语言处理中进行文本分类的常用特征，用以衡量一个词的文本区分力的强弱，正比于该词的频次，反比于包含它的文本数。