教材语言调查统计方法的新发展

——基于基础教育新课标人教版、苏教版、北师大版、语文版的比较

文 周美玲 苏新春 韩 杰 赵启文

〔摘 要〕对于任何一门科学而言，其成熟的主要标志是研究方法的科学化、系统化。在教材语言研究的过程中，开始我们多使用频次、文本数和分布率、累计频率等统计方法。由于教材语言具有基础性、有限性和有序性的特点，我们进一步提出教材语言统计的新方法，即复现调查法、使用度调查和频率差等教材语言统计方法。调查方法的发展和演进，有助于我们对教材语言的特点、性质和面貌有更进一步的认识。

〔关键词〕教材语言 语文教材 调查统计 版本比较

“教材语言指的是通过学校教育来实现教学目的，以教材为载体的语言”。[1]在对教材语言研究调查的过程中，我们的研究方法也在不断演进过程中。对于任何一门科学而言，其成熟的主要标志是研究方法的科学化、系统化。只有在不断演进的方法中，我们才有更好的工具，辅助我们对教材语言的研究有着更深入的认识。如最初我们对教材语言各要素多使用频率和累计频率等统计方法。再次我们结合使用文本数和分布率的调查方法，结合教材语言的特点，我们又使用了教材语言各要素的复现调查方法。在此基础上，我们进一步提出教材语言统计的新方法，即使用度调查和频率差等方法。

一、复现调查法

语文教材中对汉字、词汇的复现研究这方面一直比较薄弱。如李镭（2000）[2]、丁道勇（2005）[3]、苏新春（2007）[4]等。汉字，词汇在教材中的复现有着非常重要的意义。教材字词教学的安排如果能够很好地注意复现问题，将更有利于学生的记忆和识记能力，这将大大促进语文教材编写质量的提高。

我们对苏教版课文用字复现情况做了抽样调查，苏教版 18 册共有 4183 个字种。而第一学段则有字种数 1599 个。我们从常用汉字 3500 中随机抽样了 200 字作为研究对象，其中 120 字为一级常用汉字，80 字为次常用汉字，借此来考察苏教版汉字字种的复现率。下表是对苏教版汉字复现的调查样表。

表 1 苏教版复现率抽样 200 字样例

ID	字目	1 年级	2 年级	3 年级	4 年级	5 年级	6 年级	7 年级	8 年级	9 年级	总频率
23	轧										0
20	讹										0
10	缎			1							1
31	冗									1	1
18	芹				1				1		2
52	茎						1			1	2

(1) 在常用汉字中，有 22 个汉字是零出现。它们是 轧、紊、痢、痉、臼、讹、瓤、搪、恤、唁、夯、赁、泵、酗、赡、薛、碾、肄、赊、嫡、撵、遏，其中轧为一级常用汉字。

(2) 在常用汉字中，有 14 个汉字出现一次，复现

0 次。它们是：刃、缎、砾、嗜、赦、漩、讼、蛉、赘、谒、冗、垛、蝎、昭。其中为刃、缎为一级常用汉字。

⑶ 在常用汉字中，有 16 个汉字出现 2 次，复现 1 次。它们是：垄、歼、芹、嗅、唆、飒、揩、缆、鹦、涤、侥、臊、谆、猬、皿、涎。其中垄、歼、芹为一级常用汉字。

从苏教版教材抽样分析的结果来看：第一，常用汉字总体上在复现率比次常用字高，分布上比次常用汉字均匀。120 个汉字的总频率从 0~1395 次不等。80 个汉字的总频率从 0~27 次不等。第二，次常用字低频的分布基本在小学的高年级和初中。即它的初现年级是比较高的，这也就说明这些汉字在日常生活使用频率低，学习难度也大，所以教材的编写者在汉字词的学习中还是有所考虑的。

二、使用度及在教材语言统计中的使用

“使用度”是指“某调查对象的频率与分布率综合计算得出的值”。尹斌庸先生曾提出类似使用度的概念，即“通用度”。他认为：“通用度”，是指词语在语言应用的各个领域里常用性的综合指标。通用度概念中所说的“领域”，既可以指“空间”，也可以指“时间”，它既可指一个词在共时的语言应用中各领域里的通用程度，也可指一个词在历时的各个时期里的语言应用中的通用程度。[5]使用度综合考虑了频率与分布率的作用。在使用度的构成中，既有频率的因素，也有分布率的因素。《现代汉语常用字表》在对汉字统计时曾使用了使用度作为统计的计算方法；《现代汉语频率词典》大规模地运用了使用度计算法，并把使用度的计算结果作为词语选择的主要依据。

尹斌庸先生（1994）在文章中例举了词 A 和词 B 的分布和频度统计情况。如果单从分布来看，我们应该优先选词 A。但如果从两词出现的频次来看，词 A 不见得比词 B 更常用。所以单纯从分布或从频次来考查词的常用性，往往会出现较大的片面性。

表 2 两词的频次与分布数比较

	Ⅰ	Ⅱ	Ⅲ	Ⅳ	Ⅴ	频度合计	分布指数
词 A	2	1	4	1	2	10	5
词 B	4	0	7	5	8	24	4

我们在对四套语文教材的统计调查过程中也就遇到类似的问题。我们在明确到底多少词汇为教材和教学的一个基本量和常用量时，就使用两种不同的方法来调查教材的常用词汇。

表 3 四套教材课文词语分布和频次调查比较

分布调查结果			频次调查结果		
词语	课文数	频次	词语	课文数	频次
脸颊	27	27	贾芸	1	41
迷人	27	28	瑞恩	1	41
往日	27	28	杨志	1	50
朦胧	27	29	狗娃	1	52
出色	27	32	鲍西娅	2	63
送来	27	32	斑羚	1	81

从表 3 中我们看到如单从频次来看，频次调查结果一栏的的词语的频次普遍高于分布调查一栏结果的词语频次，但就词语的课文的分布的课文数来看，频次较高的课文分布却很低，只在 1~2 篇课文中出来，它们因在某一篇课文中因反复出现而进入了高频词范围。而且这些词语都是属于专名，如人名、地名、动物名等。即它们还不能算为教材中常用词汇部分。

三、频率差及在教材语言统计中的使用

我们在调查语文教材的同时也展开了对诸如历史、地理等学科教材的语言面貌的调查，那怎么可以得到几套学科教材中的特色词汇呢？我们就使用了“频率差”的教材语言调查方法。

“频率差”是指“用某调查对象在分类语料中的频率减去其在全部语料中的频率所得到的值。也叫‘频率差值’。简称‘频差’”。在词汇调查中，频率差比的是同一个词的两个频率，一个是部分频率，一个是总体频率，不属同一个总体的成员，相互之间的频率没有可比性。[6]

求频率差的步骤有四步：第一步求一个词在各分表的频次之和。第二步，求合表中所有词的总频次。第三步求合表后每个词的频率，即总体频率。第四步求每个词的各分表频率与合表频率的差值。

频率差的目的是通过观察部分频率与总体频率之间的差异来达到观察这个词在“部分”中的重要性。部分频率比之于总体频率，“顺差”愈大说明这个词在这个“部分”中愈重要、愈高频。反之亦然。换句通俗的话说，就是看一个词在总体频率中的“贡献”

如何。“贡献”大的就会在频率的比差中体现出来。频率差比的不是频率绝对值，而是二者的差异程度。如“的”字，无论是在部分频率还是总体频率中都很大，二者之间的差别很小，就很难看出它在“部分”中的特点。而“爷爷”一词，在语文教材频率高，而在各科教材中的频率不高，这就说明“爷爷”是语文教材中的高频词、特色词。

表4 语文、历史、地理三科教材词汇中最有特色的前100条词

历史教材特色词100条	工业革命、西汉、隋、北京人、废除、沙俄、镇压、版、内战、航路、北魏、罗马帝国、垄断、秦朝、项羽、变革、康有为、文艺复兴、朱元璋、集团、法制、十一届三中全会、活字、专制、巴黎公社、幕府、左图、国共、年号、义和团、资本家、反革命、北伐、苏军、原文、同盟、社会主义、印刷术、进士、东晋、李大钊、氏族、赔款、执政、中古、战败、南北朝、首脑、国民政府、主编、中国共产党、珍珠港、大败、解放区、开办、蒋介石、明治维新、晚期、影视、各级、汉人、夺取、大权、独立宣言、好莱坞、山顶洞人、戊戌变法、爵士乐、尼克松、沙皇、法典、确立、文化大革命、巴拿马、抗日救亡、手工业、齐桓公、宣言、中华民国、各组、卢梭、反帝、歼、两河流域、体制、代表作、反击、纪年、封建、议会、席位、选拔、亚太经合组织、秦始皇、北宋、隋朝、世贸组织、叛乱、中国人民志愿军、大别山
地理教材特色词100条	降水、盆地、水资源、比例尺、北半球、撒哈拉、干流、储量、铁矿、秦岭、极地、半球、经度、地下水、亚马孙河、台湾岛、复习题、中南半岛、占有量、北极圈、岛国、塔里木盆地、谷地、经纬、一般来说、气温、青海省、丰沛、地球仪、气流、高寒、进度、季风、外汇、亚欧大陆、温差、山地、农业区、西半球、巴西利亚、地形图、普查、牧区、用地、阶梯、界线、港澳、旱季、生产国、国道、冰盖、和服、水利枢纽、西北地区、地形、温带、分布、纬度、矿产、里海、大堡礁、山东省、白令海峡、沧海桑田、油棕、直辖市、大洋洲、山脉、自然资源、台湾省、旅游业、北极熊、外向型、疏松、渔场、甜菜、海拔、北冰洋、海平面、断流、侨乡、柴达木盆地、寒潮、风速、海豹、分水岭、考察站、近些年、东半球、外运、用水、用水量、地势、热带、亚热带、气候、山西省、河段、南水北调、毫米
语文教材特色词100条	听见、身子、瞧、武松、么、鸟儿、猫、妈妈、哦、燕子、那位、似的、小溪、好看、屋子、一会儿、蝴蝶、唉、一点儿、窗外、吸、坐下、妈、蜜蜂、闪、心里、摘、咱们、使劲、那儿、嗓子、跟前、松树、笑声、乡下、哇、小姑娘、灰尘、怀里、爸爸、低声、大熊猫、头上、要是、抬起、去年、并不、您、屋里、刚才、撑、身旁、舒服、底下、闪闪、鬼子、静静、平常、儿、忽然、麋鹿、呵、爷爷、脸、疲倦、那边、笑容、大哥、看见、尾巴、不曾、护士、溪流、挣、牛郎、摸、慌、安静、蒲公英、舌头、晚饭、回头、呀、清脆、匆匆、果子、波浪、村子、这儿、唱歌、歪、摇晃、哟、多久、哪怕、寂寞、焦急、力气、玩耍、曲子

四、小结

在教材语言研究的过程中，开始我们多使用频次、文本数和分布率、累计频率等统计方法。由于教材语言基础性、有限性和有序性的特点，我们对教材语言的研究进一步发展到对教材汉字词汇复现的研究。在此基础上，我们进一步提出教材语言统计的新方法，即使用度调查和频率差等教材语言统计方法。调查方法的发展和演进，有助于我们对教材语言的特点、性质更进一步的认识。各种统计方法都有各自的用武之地，词语的使用频次是最基础的数据，但就常用词的筛选来说，使用率和分布率的算法得出的结果更符合人们的语感；频率差可以凸显几套不同学科教材的特色词汇，有助于我们分析不同学科的教材语言的特色。统计方法本身无所谓优劣好坏，使用哪种方法进行计算，与研究目的直接相关，我们期待着更多更好的统计方法不断涌现。

参考文献：

[1] [4]苏新春等.教材语言的性质、特点及研究意义[J].语言文字应用，2007，(4).

[2] 李 镪.中小学语文课文字词分布统计及应用价值 [J].语言文字应用，2000，(3).

[3] 丁道勇.小学一年级语文汉字重复与识字效果关系的研究[J].课程·教材·教法，2005，(9).

[5]尹斌庸，方世增.词频统计的新概念和新方法[J].语言文字应用，1994，(2).

[6]苏新春.词汇计量及实现 [M].北京：商务印书馆，2010.

[7]周美玲，苏新春.四套基础教育语文教材的用字状况调查及思考[J].上海教育科研，2009，(4).

[8]义务教育课程标准实验教科书《语文》[M].北京：人民教育出版社.

[9]义务教育课程标准实验教科书《语文》[M].南京：江苏教育出版社.

[10]义务教育课程标准实验教科书《语文》[M].北京：北京师范大学出版社.

[11]义务教育课程标准实验教科书《语文》[M].北京：语文出版社.

〔周美玲 苏新春 厦门大学中文系 361005 韩 杰 嘉应学院宣传部 514005 赵启文 安徽省怀宁中学 246100〕