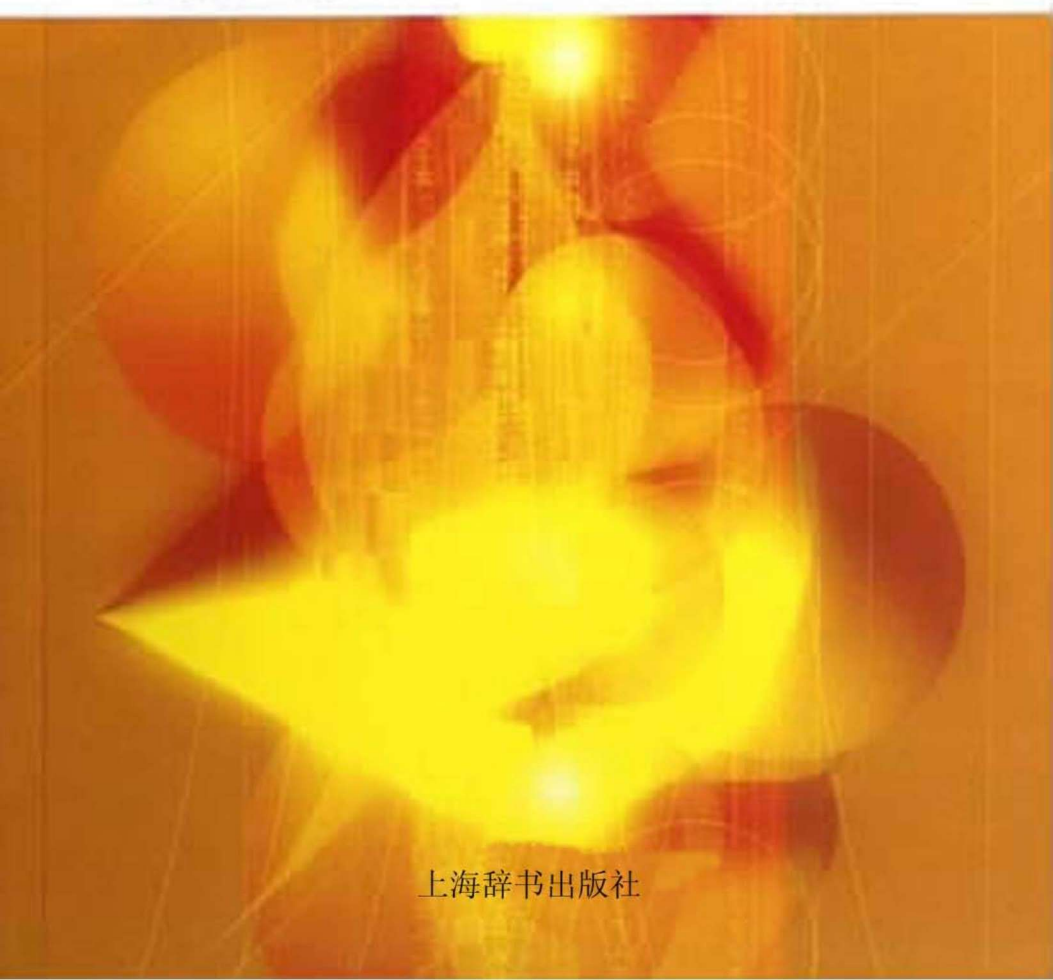


XIAOXUE JIAOXUE
JICHU HANZI DENGJI ZIBIAO

小学教学 基础汉字等级字表

陶本一 主编



上海辞书出版社

图书在版编目(CIP)数据

小学教学基础汉字等级字表/陶本一主编. —上海:
上海辞书出版社, 2012. 1
ISBN 978—7—5326—3558—0

I. ①小… II. ①陶… III. ①汉字—文字学—小学教师—师资培训—教学参考资料 IV. ①H12

中国版本图书馆 CIP 数据核字(2011)第 220232 号

小学教学基础汉字等级字表

上海世纪出版股份有限公司 出版、发行
上 海 辞 书 出 版 社
(上海市陕西北路 457 号 邮政编码 200040)
电话: 021—62472088

www.ewen.cc www.cishu.com.cn

上海展强印刷有限公司印刷

开本 890×1240 1/32 印张 12 插页 1 字数 296 000
2012 年 1 月第 1 版 2012 年 1 月第 1 次印刷
ISBN 978—7—5326—3558—0/G·773
定价: 38.00 元

如发生印刷、装订质量问题,读者可向工厂调换
联系电话: 021—66511611

序 言

陶本一教授领导的课题组,经过五年努力终于完成了《小学教学基础汉字等级字表》(以下简称“字表”)的研制工作。可以预期,这份字表将在小学教学、教材编写、辞书编纂、扫盲等方面发挥很好的作用。它为这些工作在选取基本用字方面提供了科学的依据和标准。我读过各式各样的字表、词表,也参加过一些字表、词表,如《现代汉语常用词表》等的研制,深知研制字表和词表的困难和艰辛。但眼前这份“字表”却很有特色,它在以下几个重要方面有所创新。

一、这是一份能满足现代小学教学需要和时代发展要求的字表。

社会用字是发展变化的,不同时代的字表有不同的需求和内涵,现代的字表要反映当代社会用字的现实。小学语文教学用的字表要反映小学教学用字的真实情况和需要。这份“字表”是时代应运而生的产物,是信息化时代的产物,它在前人的基础上大大前进了一步,可以说达到了历史的新水平。

过去的字表注意力集中在识字量多少上面,至于识什么字考虑就不够,更不大注意这些字在小学各年级出现的顺序,以及这些字同其他学科用字的协调等问题。本字表针对上述问题,对传统的统计

方法和统计范围进行了调整,得出了更有意义的结论。除此之外,课题组还对小学生的课外阅读进行了大规模的调查研究,注意到让“字表”同课外阅读(这是小学生语文生活的重要组成部分)协调互动。据我见到的材料,这可能是研制字表方法上的创举。

正是基于上述情况,我认为这份字表可以解决当前小学语文教学识字中的许多重要问题。首先是在组织教学中各年级不同阶段识多少字,识什么字的问题有了可靠的明确的标准;其次,解决语文学科和其他学科的用字的统筹协调问题;再次,解决更有效地指导学生进行课外阅读的问题。自然,这个字表也是我们教材编写、辞书编纂难得的参考资料,也是一般识字教育、汉字能力测试,特别是扫盲工作不可或缺的资料。

二、这份“字表”的创新性体现在哪里?

创新性是这份字表突出的特色,也是它的价值和成就所在。正是由于课题组在指导思想和研制方法上的创新,才取得了超越前人的可喜成绩。

1. “创新”首先是指导思想的创新。正如前文所说,本字表制定时就立足于小学语文教学的每个环节,要解决前人忽略或未曾解决的问题。所以“字表”的任务,不是简单解决识字量的问题,或在众多识字量差异很大的字表中找出折中的办法,研制一个字量适当的字表的问题。研制工作开始就确定不仅要解决识字量的问题,还要解决识什么字、先识什么字、后识什么字以及怎样识字等问题。他们要求每一步先调查研究,经过科学论证,再拿出结果。正因为课题组有一个统筹全局,注意各个有关方面,调动一切力量和手段来研制字表的指导思想,才能研制出一个完满地解决小学阶段全方位用字问题

的“字表”。由于指导思想先进,就站得高,看得远,所以其研究的视角和手段能够大大超越过去。比如对小学语文教材进行历时(不同时代)和共时(不同地区)的分别比对研究,对小学教学中语文学科同其他学科不同用字的全面研究,对小学生课外阅读的普查,特别是信息化时代对语言处理的先进手段的充分利用等等。这些措施保证了“字表”的科学性和先进性。

2. 小学生的语文生活,除了课堂教学外,在日常生活中与其密切相关的是课外阅读。这也是他们成长发展的重要途径。所以,这个领域中的用字也应该是小学教学需要重视的。但这个领域的用字情况如何,还没有过大规模调查研究的材料,课题组第一次这样做,必然会遇到许多困难,要花费大量的人力和物力。他们决定这样做是要有决心和远见的。事实说明,这样做的好处很多,只有包括了这方面的内容才能制定出一个小学教学阶段真实的小学生社会语文生活用字表,才能更好地改进我们的教学,提高教学的质量,解决小学生现实生活中的用字问题。自然,这也是“字表”研制中的创举。

3. 不同学科之间的大协作。课题组在研制字表时,就考虑到虽然研制的是小学语文教学的字表,但这个字表必然涉及其他学科,如其他人文学科和数学等自然学科的用字问题,而语文课也应该注意其他学科中一般用字和特殊用字的问题。所以在字表研制中,他们不仅调查其他学科用字,搜集教材,建立语料库,同时还邀请自然科学和计算统计方面的专家参与研制,听取他们的意见。正因为这样,“字表”的内容才能突破语文一科的局限,全面反映小学阶段其他人文学科及自然学科的用字情况,也才能真正成为小学语文教学用的“字表”。

4. 语料库的建设和统计方法的创新。语料库是信息化时代处理语言的必要措施和手段。但搜集什么语料,建什么样的语料库却

很有讲究。我认为在这方面课题组是做得很完备的。首先借助权威的一般语料库,了解汉字在社会中使用情况,然后进一步为自己研制的特定对象小学语文教学中的教材,建立有代表性的各种不同教材的综合语料库,这就把一般和特定的研究结合了起来。他们还进一步把小学生的社会语文生活——课外阅读的材料建成语料库,使研究的问题大大深化了。与此同时,课题组还建立了小学其他学科用字的语料库,就把研究的视野极大地扩展开了。再经过人工干预,其结果自然使得出的成果更具科学性和实用性。这比单纯从一般语言使用的语料去研究小学语文教学用字问题要全面得多、高明得多,自然也准确得多。因此,他们经过统计分析、比对研究得出的数据和结论是有坚实的语料支撑的。

特别值得我们重视的是关于每个字的使用度计算方法的改进。要真实反映一个字的使用度,除了字频之外,还有分布的问题,这是反映使用度的重要方面。但二者如何综合统一分析考虑,用科学的计算方法反映出来,我们的学者想了许多办法,使用了不少计算方式,自然各有优劣,但还不能令人完全满意。课题组经过反复研究,在前人的基础上找出了他们对特定对象(汉字)的使用度计算方法。经反复用别的手段核查,证明这种方法在目前计算字的使用度上是较科学的,准确度是很高的。这是值得我们借鉴和推广的。

这份“字表”自然不是尽善尽美的,它既可能有指导思想和方法手段的局限,也会有个别的失误,我相信在使用的实践中会进一步完善。我从阅读“字表”中学到不少东西,对我们今后研制字表有许多启发。这是我读了这份“字表”后的一点感受,不当之处请大家指教。

李行健

2011年5月11日

目 录

序言	李行健	1
小学教学基础汉字等级字表研制报告		1
小学教学基础汉字等级字表(按使用度排序)		28
一级字表 794 字		28
二级字表 901 字		61
三级字表 830 字		98
四级字表 535 字		132
小学教学基础汉字等级字表(按音序排序)		154
一级字表 794 字		154
二级字表 901 字		159
三级字表 830 字		164
四级字表 535 字		169
小学教学基础汉字等级字表(按笔画排序)		173
一级字表 794 字		173
二级字表 901 字		177
三级字表 830 字		182

四级字表 535 字	187
小学教学基础汉字等级字表(按部首排序)	191
一级字表 794 字	191
二级字表 901 字	222
三级字表 830 字	255
四级字表 535 字	285
附表	308
附表一：小学教学语料库总字频表(5 907 字)	308
附表二：小学语文教材语料库字频表(5 661 字)	324
附表三：课外读物语料库字频表(4 845 字)	339
附表四：小学其他学科教材语料库字频表(3 947 字)	352
附表五：各类专用字	363
附录	365
附录 1：语料库语料来源及取样说明	365
附录 2：小学教学基础汉字等级字表汉字部首归部说明	368
后记	372

小学教学基础汉字等级 字表研制报告^[1]

1. 研究背景

2002年4月19日《人民日报·华东新闻》第三版刊登了姜泓冰《一年级,学会读书看报》的报道,介绍了上海市二期课改的情况,称上海市将对识字教学全面提速,要求小学生在一二年级两年内认识2500个汉字,并掌握其中的1500个。同样是上海市,2006年2月10日《新闻晨报》称:本市公布《小学低年级语文教学的调整意见》为语文教材“瘦身”——识字减205个,课文删14篇,把识字教学的重心后移。识字教学这种忽增忽减、忽进忽退的情况,在建国以后的历年小学语文教学大纲或课程标准中表现得更为明显(见图1)。

1950年教学大纲规定小学阶段的识字量是3000字,1963年的大纲增加到3500字,1978年减少到2500字,2000年又回到3000字,其间变化幅度很大。低年级识字量为什么会出现这种变化?变化的依据是什么?围绕识字量和识字规律我们进行过哪些探索?上

[1] 上海市十一五哲学社会科学规划课题“基于语料库的上海市基础教育教材语言资源的建设和应用”于2006年批准立项,2008年7月研制出《小学教学基础汉字等级字表》初稿,通过国家语委语言资源监测中心的专家鉴定。2008年9月至2009年9月,在专家建议的基础上,课题组扩充语料库规模,改进等级字表的划分方法,研制出《小学教学基础汉字等级字表》二稿。

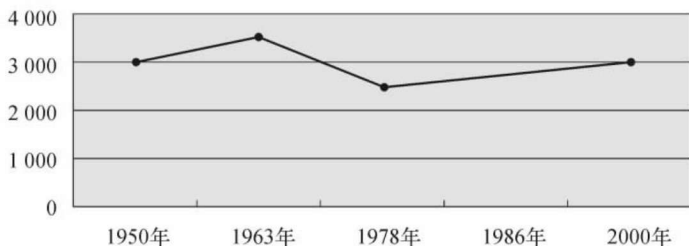


图1 历年小学语文教学大纲规定字量变化

海市二期课改语文教材主编陶本一教授认为：“现在学生要识的字是1 000个还是2 000个，为什么是这个量，它的根据是什么？还有一个认字的顺序问题，先识哪些字，后识哪些字，很有讲究。这些地方我们研究不够，严格地讲，缺乏科学依据。但是，只有解决了这些问题，才能真正提高识字效率。”^[1]

2004年起，教育部组建“国家语言资源监测与研究中心”，并于2005年和厦门大学共建了“国家语言资源监测与研究中心教育教材语言分中心”。2006年，教育教材语言分中心启动了首批科研课题，开展了大规模的教材语言研究。此后，上海师范大学陶本一和教育学院的几位教师以及研究生组建了“上海市基础教育教材语言资源的建设和应用”课题教研组，以提高对识字教学的认识，探索识字教学的科学规律。

2. 研究目标与设计

2.1 总体目标

(1) 建立3 000字左右的小学教学基础汉字表；(2) 根据汉字使用度（综合考量频次和分布状况）制定小学教学基础汉字等级字表；(3) 探讨

[1] 梁山，《在阅读中学语文》，《语言文字周报》，2005-09-06。

影响学生识字的因素,研究学生识字的规律,提高学生识字效率。

2.2 具体目标

我们要了解和研究学生的识字现状:一是了解小学阶段语文和各科的教材用字状况,二是了解学生在生活、学习中的实际用字。为此,我们从两个方面开展工作:

首先,建设小学语文教材、小学其他学科教材和课外阅读三个语料库。

(1) 建设小学语文教材语料库。选择范围立足上海,面向全国,主要收录了正在使用的上海市二期课改(教委于2000年启动)基础教育现行版语文教材(简称沪教版);为研究语文教材的演变,我们收集了上海市一期课改基础教育S版和H版语文教材(1988年开始使用);为做好同类教材的比较,我们还收集了现行江苏教育出版社(苏教版)、人民教育出版社(人教版)、北京师范大学出版社(北师大版)、语文出版社(A、B版)共八个版本的小学语文教材。

(2) 建设小学其他学科教材语料库。小学识字教学,除语文学科承担着主要任务外,其他各学科也都应该担负这方面的任务。为了解各科教材用字之间的关系,我们同时收录了除英语以外上海市二期课改基础教育各学科现行版教材^[1]。整个语料容量510余万

[1] 由于涉及版本较多,各版本教材又处于修订期,在此仅列出语料库几种主要教材的版本。现行人教版语文:一上、一下和二上为2001年版,二下为2002年版,三至六年级分别为2003、2004、2005、2006年版;现行苏教版语文:一上、二上、六上和五下为2005年版,其他各册均为2006年版;上海S版语文:一至五年级分别为1993、1994、1995、1996、1997年版;上海H版语文:一至三年级为1995年版,四年级为1996年版,五年级为1997年版;现行沪教版语文:一上、一下、四下、五上为2004年版,二上、二下、五下为2005年版,三上、三下、四上为2006年版;语文版七下、八上为2002年版,八下、九上和九下为2003年版,七上为2004年版。现行上海数学教材以2005、2006年版为主;现行上海科学、地理、信息、政治等11种教材也以2005、2006年版为主。

字。对教材作这样的选择可以实现语料的横向和纵向对比,也可以实现学科交叉对比。

(3) 建设课外阅读语料库。2006年10月至2007年5月,课题组在语文报社的支持下进行了一次全国范围的中小學生课外阅读调查。

有关课外阅读调查的问卷设计。一方面依照教育部颁布的《语文课程标准》中课外读物的建议,同时参照一些学者专家开列的推荐书目,以及主要网站发布的图书发行榜,经过反复论证、征求意见及试验性调查,确定了调查问卷。我们开列了59种图书供学生选择。同时,还设计了开放性的问卷,供学生自由填写他们喜欢的其他图书。此外我们还分别选取在学生中影响较广、社会舆论认可、发行量较大的5类26种期刊和4类13种报纸供学生选择。所选的图书、报纸、期刊既有人文学科的,也包括自然学科的。

有关课外阅读的调查对象。本次调查共涉及全国25个省市(自治区)66所学校近万名学生,还有21个省市(自治区)47所学校的近6000名家长。调查共获取16481128个有用的数据。在取样上,我们采取了分层取样与随机取样相结合的原则。根据地区经济发展不同将全国学校分成东、西、中三部分。每个省市随机选取三所初级中学,按办学层次分为:省市级重点学校、城镇级一般学校和农村学校。每个学校按每班50个左右样本进行取样,样本分布主要在初一年级段,通过对初一年级段阅读状况的了解,可知其整个小学阶段的阅读状况。

关于课外阅读调查的取样方法。根据问卷统计的结果以及对结果的专题论证,我们选取了调查结果中各排名前十位的图书、报纸和期刊作为语料的来源,建立了约110万字的中小學生课外读物语料库。

其中排在前 10 位的图书为：《安徒生童话》《格林童话》《十万个为什么》《鲁滨逊漂流记》《一千零一夜》《射雕英雄传》《哈利·波特》《马小跳系列》《伊索寓言》《钢铁是怎样炼成的》。

排在前 10 位的杂志为：《读者》《故事会》《幽默大师》《故事大王》《青年文摘》《意林》《科幻世界》《格言》《中学生阅读》《我们爱科学》。

排在前 10 位的报纸为：《语文报》《作文报》《中国电视报》《语文周报》《中国少年报》《中国中学生学习报》《体坛周报》《中国青年报》《篮球报》《中学生学习报》。

其次，在建设教材语料库和小学生课外读物语料库的基础上，通过并库研究，了解教材用字和课外读物的用字情况，建立小学生常用字表的来源字库，为等级字表的研制打下基础。

3. 研制过程

3.1 建设语料库

首先，我们建设了总规模为 6 198 337 字的三个语料库：

(1) 小学语文教材语料库。包括沪 S 版、沪 H 版、沪教版、人教版、苏教版、语文（A、B 版）、北师大版小学语文教材，共计 4 146 839 字。

(2) 小学其他各科教材语料库。包括沪教现行版小学数学、历史、品德与社会、科学、地理、美术、体育、信息科技、音乐（唱游）、自然、政治（公民）、探究性课程、争章手册（少先队活动）共 13 种教材，共计 966 839 字。

(3) 课外阅读语料库。包括各类图书、杂志、报纸等课外阅读材料，共计 1 084 659 字。

3.2 研制总字表

通过选取有代表性的语料库,能很容易统计出该语料库所含字的字频,字频在一定程度上能较好地反映字的使用程度,它的稳定性已被统计学理论和实践所证明,但字频仅能体现字的使用程度,不能反映汉字在不同语料中的分布状况。除了字频外,字在不同领域的分布状况,即所谓字的散布系数(dispersion),对衡量字的常用程度也起着关键作用。为此,我们采用使用度(usage,将字频和散布系数有机地结合起来)作为衡量汉字常用程度的指标。我们在对字表进行人工干预,以及对专家的干预结果进行统计分析时,把汉字的构词能力作为一项重要因素来考虑。

在总字表研制的过程中,我们主要按以下步骤操作:

(1) 字频统计:根据总语料库的字频统计,得出小学教学基础汉字总字频表(共有字种 5 907 个),其中语文教材共有字种 5 661 个,其他学科教材共有字种 4 845 个,课外阅读共有字种 3 947 个。

(2) 分布状况的统计:为统计汉字在不同材料中的分布状况,我们将语料库分成了 12 个子领域:沪 S 版、沪 H 版、沪教版、人教版、苏教版、语文 A 版、语文 B 版、北师大版、沪版小学其他各科教材、图书、杂志、报纸。

(3) 使用度计算:比较各种使用度公式,选择较为合理的一种进行使用度计算,计算结果按降序排列,得出汉字使用度字表。

目前国内使用度的计算公式大约有十余种:(1) 国家语委《现代汉语常用字表》使用度计算公式(国家语委汉字处,1988)^[1];(2) 常

[1] 国家语委汉字处,《现代汉语常用字表》,北京:语文出版社,1988。

宝儒汉语词汇使用度公式(常宝儒,1989)^[1]; (3) 林联合汉语词汇使用度计算公式(林联合,1986)^[2]; (4) 北京语言学院《汉语词汇的统计与分析》一书中采用的使用度公式(北京语言学院语言教学研究所,1985)^[3]; (5) 北京语言学院《现代汉语频率词典》一书采用的使用度公式(北京语言学院语言教学研究所,1986)^[4]; (6) 郑家恒的标准频率指数法(郑家恒,1990)^[5]; (7) 尹庸斌等的使用度计算公式(尹斌庸、方世增,1994)^[6]; (8) 孙茂松使用度计算公式(孙茂松,2000)^[7]; (9) 汤志祥的使用度计算公式(汤志祥,2001)^[8]。从数学统计的角度来看,我们参考了比较严谨的(6)(7)(8)种使用度公式来计算“小学教学基础汉字等级字表”中的汉字使用度。具体方法如下:

方法 1, 标准频率指数(SFI)法

散布系数 D :

$$D = \left[\log \left(\sum p_i \right) - \sum \left(p_i \log p_i \right) \right] / \left[\sum p_i \right] / \log(n). \quad (1)$$

经过调整过的字次 U :

[1] 常宝儒,《现代汉语频率词典的研制》,//陈原主编,《现代汉语定量分析》,上海:上海教育出版社,1989。

[2] 林联合,《汉语词汇统计的步骤和词汇使用度的计算》,《辞书研究》1986(4)。

[3] 北京语言学院语言教学研究所编,《汉语词汇的统计与分析》,北京:外语教学与研究出版社,1985。

[4] 北京语言学院语言教学研究所编,《现代汉语频率词典》,北京:北京语言学院出版社,1986。

[5] 郑家恒,《标准频率指数统计法在汉语词频统计中的应用》,山西大学学报(自然科学版),1990(13-2): 174-178。

[6] 尹斌庸、方世增,《词频统计的新概念和新方法》,语言文字应用,1994(2): 69-75。

[7] 孙茂松,《关于词汇使用度的初步研究》,语言文字应用,2000(1)。

[8] 汤志祥,《当代汉语词语的共时状况及其嬗变: 90年代中国大陆、香港、台湾汉语词语现状研究》,上海:复旦大学出版社,2001。

$$U = (TU/N)(FD + (1 - D)f_{\min}) \quad (2)$$

其中: $TU = \begin{cases} 10^6 & N \geq 10^6 \\ 10^5 & 10^5 \leq N \leq 10^6 \end{cases}$, N 为语料库中总字次, F 为某字的总字次, D 为散布系数, $f_{\min} = \sum ((f_i S_i)/N)$, S_i 为第 i 组语料中的总字次, f_i 为某字在第 i 组语料中的总字次(文中 $i = 1, \dots, 12$)。标准频率指数 SFI :

$$SFI = 10(\log U \div 4) \quad (3)$$

此公式中散布系数的计算采用了熵的方法,因此散布系数的计算很准确,对于字的反应很灵敏,调整后的 U 不仅参考了散布系数还参考了各组的字次情况,所以计算出的 SFI 是综合的字次。以“标准频率指数”作为统一衡量字的使用程度的标准是比较科学可行的。

方法 2, t 阶频度法

本文将抽样统计的全部语料分为 12 个子领域语料库,某一个字在 1, 2, $\dots$, 12 组的出现次数分别为 $n_1, n_2, \dots, n_{12}$, 那么这个字的 t 阶频度定义为

$$N_t = \left(\sqrt[t]{n_1} + \sqrt[t]{n_2} + \dots + \sqrt[t]{n_{12}} \right)^t / (12)^{t-1} \quad (4)$$

阶数越高,则分布对频度值的影响越大,频度值收缩越快。(这又意味着什么? 对我们的计算有什么影响?)

方法 3, 指数函数法

设语料库总字次为 N , 共分为 12 个子领域语料库, 记各子领域语料库分别含 $N_1, N_2, \dots, N_{12}$ 个字次, 并记 $N_1, N_2, \dots, N_{12}$ 中的最大者为 $N_{\max}$, 任一字 u 在相应子领域的字频分别为 $n_1, n_2, \dots, n_{12}$,

$$\text{令 } p_i = \frac{N_i}{N} \quad (i=1, \dots, 12), \quad (5)$$

$$m_i = \sqrt{\frac{N_{\max}}{N_i}} \times n_i \quad (i=1, \dots, 12), \quad (6)$$

定义任一字 u 的使用度为:

$$n_u = \mu_0 e^{-\frac{\sigma}{12\mu}} \quad (7)$$

其中: $\mu_0 = \sum_{i=1}^{12} p_i n_i$ 为字 u 的平均字频, 数学期望为 $\mu = \frac{1}{12} \sum_{i=1}^{12} m_i$, 标

准方差为 $\sigma = \sqrt{\frac{1}{12} \sum_{i=1}^{12} (m_i - \mu)^2}$, $d = e^{-\frac{\sigma}{12\mu}}$ 为 u 的散布系数, α 为常

数, 由实验测定(本文取 4), 这里指数函数 e^x 的数学性质: 当 x 趋于 0 时, 其函数值趋于 1, 当 x 趋于无穷大时, 其函数值趋于 0。当 x 处于两个极端中间时, 其函数值为光滑变化曲线, 且当 x 较小时, 变化较陡, 当 x 较大时, 变化较缓。仔细分析 μ , σ 等参数的含义, e^x 的上述特点正是散布系数所期望的。

根据上述三种计算方法, 随机抽取 34 个字, 计算出它们在语料库中按使用度排列的序号(见表 1)。

表 1 一些字的使用度序号

字	总字频	方法 1 序号	方法 2 序号	方法 3 序号	字	总字频	方法 1 序号	方法 2 序号	方法 3 序号
的	195436	1	1	1	路	5132	165	166	162
上	35764	10	10	10	欢	3703	176	187	177
小	29482	20	20	20	正	6120	185	172	179
去	14910	30	30	30	画	5475	194	175	191