陈　原　主编

现代汉语定量分析

上海教育出版社

陈　原　主编

现代汉语定量分析

上海教育出版社

现代汉语定量分析

陈　原 主编

上海教育出版社出版发行

（上海永福路 123 号）

各地新华书店经销　上海群众印刷厂印刷

开本 850×1156　1/32　印张 8.75　插页 5　字数 202,000

1989 年 12 月第 1 版　1989 年 12 月第 1 次印刷

印数 1—1,700 本

ISBN 7-5320-1470-3/H·31　定价：（特精）7.10 元

目　录

现代汉语诸要素的定量分析〔序论〕

陈　原

0 这部书——

这是一部集体著作。这是一部由一些有实际经验的语文研究者，围绕着一个主题写成的集体著作。严格地说，它还不能称为专门的学术著作，但它也不是一部这个学科的教本，更不是资料汇编。这部书忠实地记录了或者说扫描了近十年来现代汉语

诸要素定量分析的若干方面的成果，它在一定意义上充实了这一分支学科(如果可以称之为"现代汉语计量学"的话)，充实了这一薄弱环节的内容。这部集体著作也许可以称为这个语言文字应用新学科近年发展的概观。

这是一种尝试。这种尝试企图将还未系统化的分支学科，从各个不同的角度将初步探索成果公之于世——这种尝试也许对学术发展是有益的，因而会受到语文学界、教育界以及读书界的欢迎。

1 来由——

三年前的冬夜，我从海外归来，回味着在异国接触到的众多的人和书，翻看着随身带回的一些学术论著——一种愿望油然而生：我想编一套应用(实用)语言学讲座①，这套讲座应当是集体著作，参加的主要写作人员尽可能是既有专业理论修养，又有专业实践经验的研究者。这种强烈的愿望之所以产生，是因为：

——国际上凡是新的学科，特别是多科性交叉学科，在它还没有产生系统的研究专著以前，常常用集体著作的方式，把有关独立研究成果发布出来；看来这样做对学科的发展是很有利的。我手头有一大堆这样的集体著作：联邦德国版《语言控制论》(Sprachkybernetik)，是语言信息学的第一部研究著作，由十四位专家写成；荷兰版《当今的非语言交际》(Nonverbal Communication Today)，是由二十多位学者对这个既属于语言学又跨越了语言学的专题进行研究的新成果；苏联版《语言和大众传播》(Язык и массовая коммуникация)，参加写作的专门研究者超过三十人②。采取集体著作的方式有它的长处，即利用分支部门从各个角度围绕着一个研究课题进行论证，往

往能够扩大视野，达到全息扫描的目的。一个集合总比组成这个集合的各个成分之和要有力得多。也许主要就是这个原因，使集体著作成为六十年代以后国际科学出版物一种值得仿效的做法。

——日本语言学家林四郎主编一套六卷本《应用言语学讲座》③，约请我参加为其第三卷《社会言语学の探求》提供一篇关于新词语(neologism)的形成及其社会意义的研究论文。这套讲座使我对集体著作的科学论著产生了新的认识，并且有了如何进行的实际经验。

——我国语言文字的应用研究需要一种良好的环境，使各方面各层次的调查和研究成果，能够在比较短的时间内得到交流；有效的方法之一就是印行集体写成的出版物；这种著作当然也不失为鼓励青年一代研究者们在完成系统专著以前公布学术成果的一种手段。

因此，我在三年前那个冬天，拟出了《应用语言学讲座》的目录；曾经想约请这一方面的学者来研究如何实现这样一个“计划”。这个拟目只复印了几份，只有很少几位同道偶然看见(不料也被敏感的出版社编辑看到了)；步子还没迈出去，实际上我已没有可能去实现这个设想：因为不久我就不得不花了大半年时间去筹划一个大型国际会议，几乎消耗了我全部精力，所有研究工作都只能停顿下来，且不说拟议中的讲座计划了。其后是旧疾复发，我不得不躺在病床上继续消耗我剩余的精力。其后是“世事纷烦”（是纷烦而不是通常所说的纷繁），我也不得不在烦人的岗位上继续消耗我本来就不多的“余热”。但值得庆幸的是，在过去几年间，我若断若续参与了一些有关现代汉语定量分析的活动。我说“参与了”，这也许是夸大了的说法，其实我只不过接触了有关的调查研究工作，不时发表过自己的意见。这些

调查研究，就是本书所探讨的大部分或主要部分的内容。在接触当中我曾鼓励有关同志把他们的成果变成文字论述，这样，这些方面的论述自然而然成为我拟议中的讲座中的一卷。这些论文，有较深入而且颇有创见的，有研究成果极为有效而论述却平平的，也有探索得很浅，或者表达得不如人意的。集体写成的学术文集，恐怕只能如此，在这里用不上"一刀切"的方法。

这就是本书的来由。

2 主题——

这部集体著作是围绕着现代汉语诸要素进行量的测定和分析这样一个主题展开的。

从定性到定量，然后又从定量回到定性——即从量的测定结果，经过分析研究，深化对本质的理解。这也许是晚近某些学科(如果不说一切学科)的发展所经由之路。特别是近几十年信息科学体系的创立(其中包括控制论、信息论、系统论以及早些时候形成的概率论、抽样论以及其后的耗散结构理论)，高技术的导入和应用（其中包括电子信息技术以及第二次世界大战后广泛应用的电子计算机)，使语言学这样古老的学科，也逐渐注意到量的测定；即不仅着重在描述或结构分析，而且以语料(corpus 语言材料)为原料进行各种量的测定。像社会语言学这样的新兴学科，带有很浓厚的实用意义的学科，也自然而然地逐渐注意到量的测定④。对语言诸要素进行量的测定不是目的；分析这些测定数据，对语言理论提出新的观念或作出新的解释；对语言文字的实际应用作出新的设想，亦即深化定性分析，这才有利于学科的发展。比方对汉字在各种文本⑤中出现的次数进行量的测定，求得其频率，这只是一种达到目标的中间过程；根据字频数据，再利用其他制约数据或参考数据，制定常用

字表，这才完成了调查研究的一个循环；自然这只是许多循环中的一个。定性分析——量的测定——深化认识或有效应用：这就是上述循环的最简单的图式。特别是在本世纪六十年代电子技术长足发展以后，对语言诸要素的定量分析方便多了，容易多了，准确多了；这当然不能推论说在电子技术发展以前就不能进行量的测定——例如在西方世界，第一部字频统计词典是德国语言学家凯定(F. W. Kaeding)在1898年利用人工统计完成的，就连测频工作常常引用的齐普夫定率(Zipf's Law)，即有名的 $F \cdot R = C$ 也是1936年公布的，至于常被人引用的曼德布洛德(B. Mandelbrot)修正公式也是在五十年代初推导出来的⑥。至于在我国，第一个进行现代意义的字频测定，是教育家陈鹤琴在1928年完成的。他同几名助手用人工方法统计了近六十万字的语料。甚至在七十年代我国仍只用人工完成了语料为二千一百多万字的字频统计——即通常所称"748工程"。所有这些先行者的例子表明，即使在电子计算机和信息科学导入以前，对语言诸要素进行量的测定已经被认为是必要的，而且实际上证明是可能的。

一点也用不着怀疑，电子计算机开辟了语言定量研究的新时代。从前要花几倍几十倍甚至几百倍人力和时间测定某种语言要素的工作，现在利用电子计算机去做，既省时间，省人力而又能得到更为准确的数据。近年我国语言文字应用领域在短短几年间取得如此可观的成果——这些成果中相当一部分已经表述在这部论文集中——，是导入电子计算机以及其他新技术的直接结果。

本节开宗明义已指出，这部集体著作是围绕着现代汉语某几方面进行定量分析这个主题，进行探索性研究的实录；它接触到的字频、词频，也揭示了制定各种规范字表的经过(原则和实

际工作),它还深入到一些特殊领域如方言亲属关系、专名学(姓氏)的计量及分析,所有这些都统一在这个主题下面。因此,这部集体著作可以说是一种有意识编集起来的专门讲座,而不是一般性的论文集。

3 字频/词频——

统计单字在文本中出现的次数(频率),这是对所有语言进行定量分析的基本点;也可以说,字频数据是研究语言结构和语言应用的基础。所以各种语言文字的量的测定总是以此为出发点的。汉语的一个特点是字和词不是任何时候都一致的——一个方块字或者称之为一个"字符",可能是一个有独立完整语义的词,也可能只是一个词的构成部分(词素);所以对现代汉语量的测定,同时要有字频和词频两种数据,缺少其中一种就不完全,不能据此论述现代汉语的全貌。

本卷头五篇论文所统计和分析的正是字频和词频这两个(一个)基本点,以及由此产生极有社会效益的常用字/词表。五篇论文的作者都分别参与了四个不同的实际测量工作,这项工作或者可以简称为"语言工程"⑦,表述并分析了四项统计结果。这四项成果都分别印制专门的数据集,可供各方面利用。这五篇论文简明扼要地提供了进行这几项工程的方法和程序,并进而就所获得的字频、词频数据做了初步分析。语文学者当然可以根据所得数据做其他有专门目的的分析,这里提供的只是一般性的定量分析。

把《现代汉语频率词典的研制》一文放在这一组论文之首,因为它的论述不只提供了编制汉语词频词典的研制过程(原则和方法),而且阐述了字频测定和词频测定的一般性原理——本文提出并且回答了样本数量的"最佳"选择问题,通俗地说,即在

进行字频词频测定选取的语料究竟达到多大的数量为最优的问题。语料数量过少，统计结果不能符合语言应用的客观实际，这是可以想象到的；而语料数量又不能扩大到无穷，数量过多，费时失事。能不能说数量越大越好呢？这篇论文根据概率论的大数定理，认为常用字词出现频率不低于 10^{-5}（即在十万次场合至少有一次出现机会）为适度的，为此，还可以增加一个数量级，即在一百万次语料中出现一次为适度。本文所述的这一项语言工程又增加了"保险"系数，实际取样为 200 万字符（不算标点符号为 181 万字符，131 万词次），以此来统计单个汉字出现的次数（频率）以便进行选定常用词，作者认为是可取的。用这种规模测量的结果是：1000 个高频单字，覆盖了所用语料的 91.3%；8000个高频词，覆盖了所用语料的 95%。这里给出的几个数字对于研究并制定常用字表和常用词表是很有参考价值的。

顺便说一下，对语言要素进行量的测定，语料数量超过了必要的最优值，那可能导致浪费。换句话说，所用语料适度就可以得出可靠的结果。例如测定现代汉语的平均信息量（熵）时，冯志伟[6]采用了逐渐增大汉字容量的方法，计算出当汉语书面语句中的汉字容量扩大到 12370 个单字时，包含在一个汉字中的平均信息量（熵）为 9.65 比特——如果汉字容量继续增大，所求得的熵值不会增加。熵和字频当然不是一码事，这里只是顺便说明，测量用的语料数量应求得最优量。

对字频测定所用语料的最优量，目前还有不同的意见。从实际的几个语言工程看来，样本数量远比本文所提出的最优量为大。试与英语字频测量比较一下。最近一次英语词频测量（1971年）用了 5088721 个字的语料，共出现 86741 个单词（即单字）。我国七十年代中期"748 工程"用人工进行现代汉语字频测定用了 21629372 个字符，而字种（对应于上例中的单词）只

有6374个；而另一个工程，即本书所论述的那一次工程，用计算机进行现代汉语字频测定，则只用了一半数量，即11873 029个字符，共得字种7745个。

本书各篇在阐述现代汉语频率测量及分析时，采用了如下的术语：

(1) 语料(corpus)——所用的语言材料，即印出的文字材料，以字数(词数)为单位；

(2) 样本(text)——有时同"语料"同义，有时指抽样用的特定文本；

(3) 字符(token)——指在语料中出现的总字数(总词数)，包括重复出现的字数(词数)；

(4) 字种(type)——指在语料中出现的不同的单字，文中也有作"不同词"的。

(5) 频率[频度](frequency)——单字在文本中出现的次数与所用语料所含总字数之比。

对汉语的定量分析必须分别处理字(方块字)和词，这是由汉语的特别构造决定的，如上所述，汉语的字有时是词，有时只是词素。因此，有字频(frequency of characters)和词频(frequency of words)之分，接着即有常用字和常用词之分，定量分析时，必须绝对区别这两者；本书头一篇论文正好把测量字频和词频的不同点扼要阐明了。作者说：

> "统计汉字的频度，有一个字算一个字，不存在词语单位的切分问题。使用拼音文字的外语，单词之间有空白间隔，统计词数也不太困难。统计汉语词频则难度要大得多。无明显形态界限作为划分词的依据，这是主要困难。语素和词，词和词组的界限划分以及词的分类问题在理论上和实践上都尚未妥善解决。"

在同一组论文中，还有一篇阐述另外一个语言工程（词频测量工程），对“词”的定义和划分方法，提出了另外一种见解，并根据这种观点利用电子计算机进行自动切分——关于这个问题的争议，留待下文讨论。

现在回到《现代汉语频率词典》所取得的几个关键性数据。根据测量结果，在这项工程选择的“最优”语料数量范围内，共测得4574个字种。在最优量语料（如前所述，约200万字符）中出现245次以上的一千个高频汉字，覆盖面（即占全部语料字符的百分比）达91.3%；如果把出现30次以上的2418个高频及次高频汉字测算，则覆盖面达到99%强。在整个语料中出现的4574个字种中，减去这部分高频和次高频汉字（即4574－2418），得2156个字种——这二千多个低频汉字只覆盖全部语料的1%。论文认为这个部分的汉字（低频汉字）每一个出现的平均机会只有千万分之五（5/10000000）。

论文对1000个高频汉字进行的语音分析和语义分析是饶有兴味的，其结果对于应用语言学，语言教育学，社会语言学，心理语言学，语言信息学以及其他学科都有启发性的意义。

语音分析的结果提供了这样的一个事实，即以Z和S子音开始的汉字（这里用的当然是汉语拼音方案）占绝对优势，仅“shi”这个音节在1000个高频汉字中即占有24个（2.4%）。为此，文中引用著名语言学家赵元任“编造”过一个《施氏食狮史》的绕口令，就是从这样的事实出发的。《施氏食狮史》从旁证明在现代汉语口语里头，复音词出现较多，不常发生因使用同音词而语义不清的情况；但如古文（文言文）今读（注意：以今音来读古文），则因为单音词多，使语义分辨发生困难。这个极端例子见于赵氏关于语言问题的演讲录（《语言问题》第十讲，《语言跟文字》），如果用汉语拼音转写，即使加注调号，读来也是颇为费

解的。这篇拗口令原文如下：

> 石室诗士施氏，嗜狮，誓食十狮。氏时时适市视狮。十时，适十狮适市。是时，适施氏适市。氏视是十狮，恃矢势，使是十狮逝世。氏拾是十狮尸，适石室。石室湿，氏使侍拭石室。石室拭，氏始试食是十狮尸。食时，始识是十狮尸，实十石狮尸。试释是事。

对1000个高频汉字进行语义分析时，作者提出了汉字的构词能力(外国有些学者称为"词力" word power)问题。现代汉字在执行交际功能时最本质的属性是它的构词能力，而过去很少对词力进行定量分析，正是这个语言工程，弥补了这样的一个极有意义的空白。测量结果是：构词能力在100条以上，出现字次在1000以上共有70个汉字——这70个汉字在这项语言工程中构成了所列词条11133条之多，占35.7%。在这七十个字中占头十个的是"子、不、大、心、人、一、头、气、无、水"。

在确定一个字或一个词是否是常用字或常用词，不能单纯依靠频率，这是很容易理解的。在进行语言定量分析，特别是进行常用字常用词测定时，要考虑到字/词的分布状态；因此导入了"使用度"(usage)这样的观念——这个语言工程推导了现代汉语词的使用度公式（后来在制订现代汉语常用字表这项语言工程中也试着推导一个使用度公式)。在现代汉语定量分析工作中，这是有重大意义的实验。

另外一个语言工程，即本组论文第二篇《现代汉语字频测定数据及分析》所论述的一项工程，也是在1985年完成的。这个工程可以说是"748工程"(用人工进行的大规模字频测定)的继续，它所用的测定样本共11873029字符（比"748工程"少一半)，论文说这是从1977—1982年问世的社会科学和自然科学文献一亿三千八百万字(138000000)中抽出的样本。遗憾的是，

不论是这篇论文，还是别的有关论文，都没有对所选样本作过详细分析；例如这里只提到样本分为四个方面（报刊、教材、专著、通俗读物）以及每个方面下面分成若干类别，这是很不够的。数量这样巨大的语料（超过一亿汉字字符，或者说，五千万上下独立词），当然不是随意选样的；这项语言工程和其他语言工程一样，在进行之初即由专家组根据一定的原则选定样本。我认为在公布每一项语言工程的全部资料时，应当首先发表全部专家组讨论选样原则和在实际上如何选样的系统意见或不同意见，然后附列选样目录。样本在定量分析中有重大意义，甚至可以说有着决定意义。看来，所抽取的一千多万字样本（如"748 工程"所抽取的二千多万字样本一样），都是全部输入计算机加以统计的。这当然是一种方法；其实也可以考虑减少样本数量，对每一个样本采取等距离抽样——例如"美国传统中级语料字（词）频统计"（AHI Corpus）即采取这样的方法，对每个选出的样本抽取其最初 500 字（如句子未完，抽到句子完了为止）输入计算机，这项工程在确定抽取每个样本最初 500 字为最优值之前，曾作过包括 100000 字符的抽样试验，每个样本抽取最初 500 个字为一组，抽取最初 2000 个字为另一组，结果认为每个样本抽取最初 500 字已可以给出"适当的弹性"（adequate flexibility），这项工程还推导了各类语料应抽取多少种样本，每种样本应抽取多少文本的公式。[9] 对现代汉语的测量，将来可以参考这些数据推导出自己的公式。

《现代汉语字频测定及分析》这篇论文写得简明扼要，提供了该项语言工程的基本数据，同时也对两项先前进行的字频测定工程进行了比较分析；此外还对分布度作了阐述。尽管"748 工程"是在特殊语境（"文化大革命"后期）下用人工方法测量的，但是它与这一次在普通语境（1977—1982）下使用电子计算机测

量的结果很相似，这从两项工程的字频曲线图可以看得很清楚。两条曲线所用的最基本数字是：

	语　料	所得字种数
“748 工程”	21629372	6374
85 年字频测定	11873029	7745

论文作者对两项字频测定工程对比研究后，提出了这样的论点：两者的函数曲线大致相仿，只是“748 工程”的曲线在字序 3000 以前略高于其他一个工程的字频函数曲线。两条函数曲线在字序号 3000 处相交，而 3000 号以后的点列极其相似。

检验两项工程的字频数据还可以发现，按降频序到 161，162 号时，两者覆盖全部语料（尽管两项工程的语料数量不同）都同时达到 50%（“748 工程”162 号为 49.97%，后者 161 号为 49.93%）。也就是说，现代汉语中使用频率最高的 161-162 个字，在实际应用中已覆盖了文本的一半——但这决不意味着掌握这 161/162 个汉字便可以了解文本语义的半数，因为理解语义这个问题比较复杂，字和词、词与词组、上下文等等，都会对理解度发生不同程度的影响。

这篇论文揭示了这么一个例子：序号为 1 的汉字（两项结果都是“的”字）的出现次数并不随着样本容量的增大而持续增大。此外，样本容量的增大并不意味着常用汉字出现次数按比例增加。某字在一千万字样本中出现一次，在二千万字样本中不一定出现二次。这项研究也同上一篇论文一样，也注意到分布率，推导了一个分布公式。尽管两个公式不完全相同，我认为将来可以通过无数次的实践加以检验和修正。作者指出：“如果今后的汉字频度统计将把汉字的分布篇数这个数据统计上，综合汉字的频率、分布类数和分布篇数这三方面的因素，就有可能对汉字作出更加准确的描述”。