**九十年代汉语词汇地域分布的定量研究[[1]](#footnote-1)**

汤志祥

1. 导言
   1. 二十世纪八、九十年代是华人社会（涵盖中国大陆、台湾、香港等地区）

在政治、经济、文化、生活等各方面都飞速发展变化的时期。这世纪之交的最

后十来年间, 在这个也叫着“汉语文化圈”地域中，凡是可以觉察到的时代演

进的每一步伐和各自社会的点滴变化都充分反映在通行的汉语 — 这一“信息

载体”上。

而语言的“四大要素” — 语音、词汇、句法、语义中，唯有词汇和语义

具备了折射时代以及反映社会的功能。这两者都是一个“动态”的体系。它们

一直是处于一种与社会同步发展变化的状态之中。从这一点出发，可以说，研

究一种语言的词汇及其变化是研究考察这个社会变迁的重要方面。反之，研究

一个社会也可以从词汇及语义的角度出发，去寻找确凿的印证。

由于历史的原因，长期以来华人社会有所谓的“两岸三地”[[2]](#footnote-2)之说。这“两

岸三地”所使用的语言是“同源、同体”的民族共同语 — 汉语。但又具有各

自不同的地域特点。因此今天研究汉语词汇，应该具有更高的“视点”和更广

的“视野” 。

0．2 当代汉语词汇研究应当具有以下五方面的特质：

1. 广泛的地域分布： 涵盖中国大陆、台湾、港澳等地区；

（2） 崭新的时期分段： 截取二十世纪八九十年代的时段；

（3） 坚实的语料基础： 运用中大型的精语料库；

（4） 客观的分析方法： 从定量到定性的分析过程；

（5） 现代化的研究手段： 电脑检索、统计和分析。

* 1. 前后历时六载，于一九九七年六月完成了全部语料录入，词语切分，并具

备了各项基本检索功能的中型当代汉语语料库 — 香港理工大学的《中港台汉

语词库》（原暂名,以下简称《汉库》）[[3]](#footnote-3)基本具备了以上五方面特质。其基本情

况和数据如下：

建库单位： 香港理工大学中文及双语学系

建库时间： 1991 年 — 1997 年

语库容量： 6,100,194个符号

汉字： 5,139,920个 （+ 960,274个非汉语符号）

词条[[4]](#footnote-4)： 60,811个 （汉语）

6,373个 （英语）

词次： 3,273,760个 （+ 218,528个专名）

收词地域： 大陆、台湾、香港三个区域

语料时段： 1990年 — 1992年

语料来源： 1、大陆： 《人民日报》、《北京晚报》（北京）、

《新民晚报》（上海）、

《羊城晚报》（广州）

2、台湾： 《中央日报》、《中国时报》、《联合报》

3、香港： 《明报》、《信报》、《成报》

编列方式： 词频排序， 普通话和粤语音序索引

编排内容： 1、词序 2、词语 3、汉语拼音 4、粤语拼音

5、词频 6、累计词频（覆盖率） 7、总使用度

8、地域使用度：（1）大陆 （2）台湾 （3）香港

9、学科使用度：（1）政治 （2）民生 / 治安 / 环境

（3）财经 （4）体育 （5）文化活动

（6）娱乐 （7）读者来信 （8）广告

（9）副刊：旅游、影评、述评、小说、

人物、艺术、科学、健康等

据悉，这是迄今为止已经完成的语料地域分布涵盖整个汉语文化圈的第一

个汉语语料库。

* 1. 《汉库》由以地域划分的大陆、台湾、香港三个独立的“分语料库”（以

下简称“京、台、港” [[5]](#footnote-5) ）的语料组成，然后再合成一个“整语料库” 。

《汉库》中京、台、港三个分语料库的语料容量数据是：

《中港台汉语词库》中三个分语料库的语料容量（字数）

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 大陆地区 | | 台湾地区 | | 香港地区 | |
| 语库量 | 比例 | 语库量 | 比例 | 语库量 | 比例 |
| 1,776,870 | 34.57 %[[6]](#footnote-6) | 1,675,614 | 32.60 % | 1,687,436 | 32.83 % |

以上京、台、港三个区域的语料量相对比例大约为：34 : 33 : 33 。即三

者的相对比例大致为：1 : 1: 1 。比例大致是相当的。

为此我们可以从横和纵的两个角度，在“整语料库”中统计和分析出整个

京、台、港三地的词语共时地域分布状况。从而从定量的考察中得到比较客观

的定性结论。

1. 当代汉语词语的地域分布
2. 1 国内汉语研究和教育界早就注意到汉语文化圈内三个不同区域之间的词语

差异。近十年来已经有不少专论和辞书专门描写和分析其中的一些差异。其中

有些是双区域的（包括大陆与台湾的“两岸”之间的词语差异[[7]](#footnote-7) 或者国内与香

港的“两地”之间）词语差异[[8]](#footnote-8) 。也有的是三区域的（即所谓“两岸三地”

之间）词语差异[[9]](#footnote-9)。

应该说，这些研究成果都是很有价值的，但是相对而言都尚有一些需要商

榷的地方。比如说：1、语料的周遍性，2、语料的代表性，3、词语的稳固性，

4、词语的出现频率，5、词语大类的外延和划界，6、词语的内涵，等等。

* 1. 从《汉库》的语料来考察，如果从地域的角度去看，京、台、港三个区域

各自使用的汉语词汇可以视作下列三个语言圈（圈子大小并不意味着词语数量

的多寡）：

而在实际的语言使用中，上述三个语言圈子是彼此相互交叠的。情形就

像下图：（外延大小并不意味着词语数量的多寡）：

A AB B

ABC

AC BC

C

图中，三个圆圈重叠的 ABC 部分代表着京、台、港三地共用，即在整个

使用汉语的地区（三个区域）都通行的部分。这部分可以看作是京、台、港三

地汉语词语相同的部分，叫做“三区域共用词语” ；

图中两个圆圈重叠的AB，BC 和AC 三个部分分别代表在其中某两个区

域（AB：京、台，BC：台、港，AC：京、港）里通行的部分。叫做“双区域

通用词语”[[10]](#footnote-10) ；

而其余完全不重叠的A，B，C 部分则各自代表着仅在其中某一个区域里

通行的词语。它们是京、台、港三地之中各自独特的部分，叫做“单区域独用

词语” 。

“三区域共用词语”（以下简称“共用词语” ）应该认为是当代汉语词语

的“共同底层” ，是华人社会信息交流底部基础；而“双区域通用词语”（以

下简称“双区词语” ）应是“共同底层”的外延，是“共用词语” 的直接补

充部分。三个“单区域独同词语” （以下简称“单区词语” ）均属于“各自

表层”部分。它们是“共用词语”的预备补充部分。

当然“双区词语” 中“京、台通用词语”和“京、港通用词语”要比“台、

港通用词语” 离“共同底层”的距离更近，进入“共同底层” 的速度也更快。

而“单区词语” 中“大陆独用词语” 也比“台湾独用词语”及“香港独用词

语” 进入“共同底层”的速度为快。其原因是： 1、大陆汉语是汉语的“母

体” ，台湾和香港汉语是“子体” ； 2、无论是使用地域的面积和使用的人

口数量， 大陆汉语都占了绝对优势；3、就单语使用普通话而言，其比例也是

前者远远高于后两者。（后两者都是“双语区”和“多语区” ）。

2． “共用词语” 、“双区词语”和“单区词语”例词

2．1 “三区域共用词语”举例[[11]](#footnote-11)

“共用词语”几乎包括所有的单音节词语和四音节词语。而且词语的序号都

比较靠前。

（1） 单音节词（按频率递降序的前五十个，序号从1 号到50号）

的、在、一、是、有、不、了、十、和、人、上、年、为、个、者、

之、他、队、将、与、到、中、而、也、及、以、说、三、多、会、

大、二、已、月、五、但、两、来、时、就、日、并、白、国、四、

于、后、我、等、能；

（2） 双音节词（按频率递降序的前五十个，序号从53号到209号）

经济、公司、政府、问题、表示、他们、国家、我们、发展、市场、

工作、没有、今年、社会、投资、认为、目前、企业、世界、进行、

可以、国际、由于、大陆、可能、有关、去年、方面、计划、第一、

一些、自己、要求、因为、指出、总统、代表、部分、改革、活动、

这些、其他、会议、地区、因此、情况、人士、举行、关系、人员、

主要、这个、时间、人民、昨天、同事、影响、希望、服务、包括；

（3） 三音节词（按频率递降序的前五十个，序号从297号到2819号）

委员会、共和国、电视台、发言人、一方面、候选人、大多数、

负责人、青少年、运动员、锦标赛、平方米、越来越、办公室、

现代化、房地产、办事处、艺术家、秘书长、不得不、可能性、

反对党、一系列、股份制、博物馆、研究所、财政部、总书记、

录影带、记者会、愈来愈、经济部、电视剧、检察官、分析家、

公务员、派出所、小朋友、爱滋病、大学生、百分点、反对派、

高科技、居英权、建筑物、自行车、进出口、国内外、出版社、

电视机；

（4）四音节词（按频率递降序的前五十个，序号从1822 号到9916号）

卡拉OK[[12]](#footnote-12)、经济学家、引人注目、平方公里、前所未有、大专院校、

管弦乐队、供不应求、众所周知、成千上万、讨价还价、举足轻重、

丰富多彩、实事求是、二氧化碳、不可或缺、千方百计、各行各业、

当务之急、农副产品、不约而同、层出不穷、脱颖而出、全心全意、

各式各样、不以为然、理所当然、意想不到、不可思议、全力以赴、

千家万户、房地产业、不知不觉、难得一见、显而易见、后顾之忧、

脍炙人口、名副其实、耳目一新、有目共睹、息息相关、安居乐业、

应有尽有、方兴未艾、刮目相看、出乎意料、不遗余力、出人意料、

相提并论、独一无二；

2．2 “双区域通用词语”举例

“双区词语”又分作三部分：（1）京、台通用词语，（2）台、港通用词

语，（3）京、港通用词语。这类词语主要是双音节词和三音节词。它们出现

的序号一般都比较靠后。现将各自的四十个例词分别例举如下[[13]](#footnote-13)：

1. 2．1 京、台通用词语
2. 单双音节词（按频率递降序，序号从3757号到6685号）

栋、 坑、县府、额度、违章、组建、摊贩、编列、违规、片子、

酱油、缓解、民营、交割、路子、公交、搭配、评比、管线、告发；

1. 三音节词语（按频率递降序，序号从4258号到12939号）

废弃物、被害人、乡公所、复印机、企事业、电信局、大奖赛、

全方位、侦查员、海洛因、合作社、交换机、学杂费、冰淇淋、

社会化、政策性、阶段性、根据地、管委会、认购证；

* + 1. 台、港通用词语

1. 双音节词（按频率递降序，序号从2863号到6629号）

民运、飞弹、疲弱、片商、私校、录影、同业、权证、预估、私家、

社工、籍着、相较、房车、核武、知会、影带、厘定、国父、执业；

（2） 三音节词语（按频率递降序，序号从5348号到13038号）

嘉年华、高峰会、影印机、偷渡客、六合彩、年增率、精神科、

大陆客、停车位、荷尔蒙、公民权、争议性、水族馆、剧情片、

生育率、民营化、同意权、移民潮、海洛英、伴唱带；

* + 1. 京、港通用词语[[14]](#footnote-14)

1. 双音节词（按频率递降序，序号从1104号到10080号）

靓、 碟、今次、通胀、船民、展销、楼宇、的士、物业、软件、

录像、弱智、汇报、首期、型号、客商、车匪、打印、拚搏、街坊；

1. 三音节词语（按频率递降序，序号从4799号到18802号）

录像带、录像机、游戏机、大耳窿、公积金、打印机、商品房、

紧迫感、大排档、直通车、赞助商、国庆节、咨委会、再生产、

朱古力、外省市、一次过、集团化、发行师、回乡证；

2．3 “单区域独用词语”举例

“单区词语”又分作三部分：（1）大陆独用词语，（2）台湾独用词语，

1. 香港独用词语。一般也是双音节词和三音节词。它们出现的序号一般都

很后。现将各自的四十个例词分别例举如下：

1. 3．1 大陆独用词语
2. 双音节词（按频率递降序，序号从3745号到10585号）

案犯、团伙、民警、房改、荧屏、粮店、房管、面料、解困、民办、

公房、评委、十佳、剧组、老伴、老区、信访、上岗、三产、动迁；

1. 三音节词（按频率递降序，序号从6418号到14479号）

居委会、供销所、文化站、监督局、粮食局、离退休、粮管所、

群众性、小商品、面包车、煤气灶、特困户、人贩子、节假日、

小公共、文化馆、豆制品、自觉性、影剧院、指战员；

2．3．2 台湾独用词语

（1） 双音节词（按频率递降序，序号从2002号到7967号）

国中、国小、安打、职棒、联考、行库、国代、课征、窃盗、国协、

软体、上档、电玩、港剧、列管、宣导、费率、查报、税捐、房贷；

1. 三音节词（按频率递降序，序号从6911号到13120号）

选委会、证交税、公权力、国中生、立法院、歌仔戏、交流道、

零组件、公交法、自营商、公私立、原住民、督察室、幻像机、

原委会、双年展、电影处、防火巷、印表机、参与率；

2．3．3 香港独用词语

（1） 双音节词（按频率递降序，序号从4063号到8555号）

按揭、视乎、公屋、求证、轻铁、中六、经已、若然、规例、楼盘、

加幅、押后、羁留、发牌、失车、沽盘、收生、楼市、学额、美仙；

（2） 三音节词（按频率递降序，序号从5741号到12201号）

大律师、开幕礼、入境处、保护令、消防处、临屋区、发展商、

颁奖礼、电话卡、平方口尺、蓝筹股、回报率、区局节、律师行、

音统处、护卫员、反黑组、民主派、评议会、赔偿额。

3． 当代汉语三个区域词汇共时分布的数量和比例

3．1 不考虑词频的纯语料库数量统计和“人工干预”[[15]](#footnote-15) 统计

如果不考虑词频，仅仅根据《汉库》纯语料库语料的统计，我们可以得出一个“语库统计”的数量和比例。如果再经过“人工干预” 后的“定性”再“定量”的计算，可以得出一个“人工干预”后的统计结果。

“共用词语”词条与总词语词条的数量及比例的结果分别如下：

《汉库》中总词语词条与“共用词语”词条的数量及比例

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 纯语料库统计 | | 人工干预的统计 | |
| 数量 | 比例 | 数量 | 比例 |
| 总词语词条 | 60，811个 | 100.00 % | 60，811个 | 100.00 % |
| 共用词语词条 | 16，742个 | 27.53 % | 55，412个 | 91.12 % |

“双区词语”的“纯语料库统计” 以及“人工干预”后的数量和比例，分

别如下：

《汉库》中总词语词条与“双区词语”词条的数量及比例

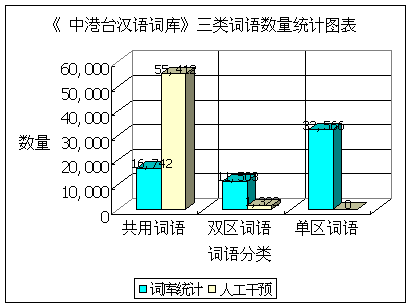
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 纯语料库统计 | | 人工干预的统计 | |
| 数量 | 比例 | 数量 | 比例 |
| 总词语词条 | 60，811个 | 100.00 % | 60，811个 | 100.00 % |
| 双区词语词条 | 11，503个 | 18.92 % | 1,322个 | 2.17 % |

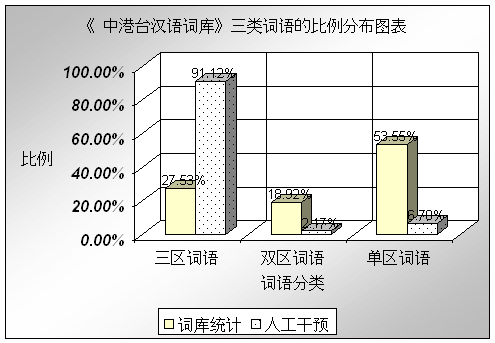
“单区词语” 的“纯语料库统计”以及“人工干预”后的数量和比例，

分别如下：

《汉库》中总词语词条与“单区词语”词条的数量及比例

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 纯语料库统计 | | 人工干预的统计 | |
| 数量 | 比例 | 数量 | 比例 |
| 总词语词条 | 60，811个 | 100.00 % | 60，811个 | 100.00 % |
| 单区词语词条 | 32，566个 | 53.55 % | 4，077个 | 6.70 % |





很明显，以上统计很清晰地反映了一个强烈的反差：根据纯语料库语料统

计，三地“单区词语”占的比重超过一半（ 53.55 % ），“共用词语” 次之（ 27.53

%），而“双区词语”最少（ 18.92 % ）。三者相对比例约为： 5 ： 3 ： 2。

但是经过“人工干预”的“定性”结果，则极不相同。“单区词语”只有6.70 % ，

“双区词语” 只有2.17 %，绝大多数是“共用词语” ，占了91.12 % 。这个

对比充分说明了两点：

（1）《汉库》的语料量对正确显示三个地区的词语的“共同底层词汇量”及

其相互之间的“差异量” 还显得不够，因此偏差还比较大[[16]](#footnote-16)。

（2）光看词语数词的绝对数量，还不能揭示汉语词汇的真实共时状况。“词

频” 、“累计词频” 、“地区分布比”和“地区词频”是考察词语分布状况的

重要参量。

3．2 考虑词频的“共用词语” 、“双区词语”和“单区词语”统计

词频是词语研究的极端重要的参量。为研究和表述方便，我们把《汉库》 的

词频的分级分成下列三段，六级：

|  |  |  |  |
| --- | --- | --- | --- |
| 段 别 | 级 别 | 名 称 | 频 率 [[17]](#footnote-17) |
| 高频词 | 第一级 | 高高频 | F > = 30,000 次 |
| 第二级 | 低高频 | 30,000次 > F > = 3,000次 |
| 中频词 | 第三级 | 高中频 | 3,000次 > F > = 300次 |
| 第四级 | 低中频 | 300次 > F > = 30次 |
| 低频词 | 第五级 | 高低频 | 30次 > F > = 15次 |
| 第六级 | 低低频 | F < 15次 |

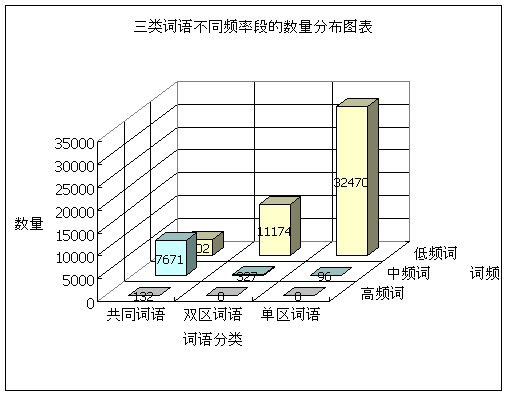
根据上述的频率表，我们再对纯语料库统计的“共用词语” 、“双区词语”

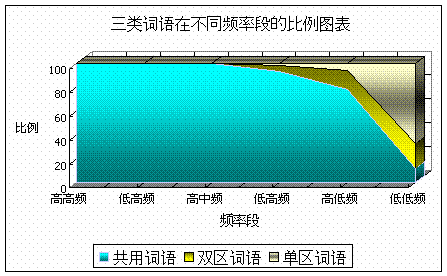
和“单区词语”的词语数量做一次新的统计。得到的数据如下：

“共用词语” 、“双区词语”和“单区词语”的词语在不同频率段的

数量和比例

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 段 级 | 共用词语 | | 双区词语 | | 单区词语 | | 总 计 |
|  |  | 数量 | 比例 | 数量 | 比例 | 数量 | 比例 | 数量 |
| 高频词 | 高高频 | 4 | 100.0 % | — | — | — | — | 4 |
| 低高频 | 128 | 100.0 % | — | — | — | — | 128 |
| 中频词 | 高中频 | 1,323 | 99.9 % | 2 | 0.1 % | — | — | 1,325 |
| 低中频 | 6,216 | 93.6 % | 329 | 4.9 % | 96 | 1.5 % | 6,639 |
| 低频词 | 高低频 | 3,502 | 78.5 % | 700 | 15.7 % | 260 | 5.8 % | 4,462 |
| 低低频 | 5,569 | 11.5 % | 10,474 | 21.7 % | 32，210 | 66.8 % | 48,253 |
| 总计 |  | 16,742 | 27.5 % | 11,503 | 18.9 % | 32，566 | 53.5 % | 60，811 |





以上的统计给了我们一个全新的概念，如果把频率考虑进去的话，就不难

发现，纯语料库统计的“共用词语”在自“高高频”到“高低频”的各个频率

段的使用频率和比例都高于同段级的“双区词语”和“单区词语” ， 尤其在

高频词的段级，百分之百是共用词语” ；即使在中频词的段级（共7,539个

词） ，“共用词语”的数量比例高达93.6 %，而“双区词语”和“单区词语”

的比例分别为4.9 % 和1.5 %。远远高于后两者；甚至到了低频词段级的“高

低频” 段级（共 3,502 个词），“共用词语”的数量也在78 % 以上。

情况仅仅只在“低低频”段级才变得相反。这样就显示出，在整个词库的

词语中，就词频而言，在“高低频”以上的五个段级中（前11,173个词），

三者排序都是：第一“共用词语” ，第二“单区词语” ，第三“双区词语” 。

3．3 考虑累计词频的“共用词语” 、“双区词语”和“单区词语”统计

“累计词频”是词语研究的又一个重要参量。它能利用词语使用频率累计的

而形成的“覆盖面”去揭示出词语使用的等级：（1）最常用词、（2）次常用词、

（3）常用词、（4）通用词。

根据国家对外汉语教学领导小组办公室汉语水平考试部，综合研究了自

1959年至1991年32年间国内主要的16种常用词统计资料后所制定及颁布的

《汉语水平词汇与汉字等级大纲》[[18]](#footnote-18) ，上述四级词语等级的划分线是：

最常用词、次常用词、常用词、通用词四级词语等级划分标准

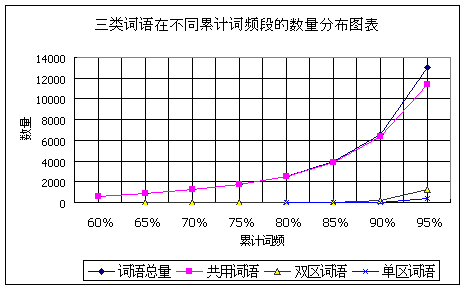
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 等 级 | 级 别 | 界 标 | 覆盖面 | 实际数量[[19]](#footnote-19) |
| 最常用词 | 甲级词 | 1，000词次 | 73 % [[20]](#footnote-20) | 1，027个 |
| 次常用词 | 乙级词 | 3，000词词 | 86 % | 2，012个 |
| 常用词 | 丙级词 | 5，000词次 | 91 % | 2，114个 |
| 通用词 | 丁级词 | 8，000词词 | 95 % | 3，907个 |
| 共 计 |  |  |  | 9，060个 |

根据《汉库》的语料统计，我们再从“覆盖面”去考察这四类词语的分级、

分等情况的话会看到如下不同结果：

《汉库》词语在不同累计词频段的分布情况

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 累计  词频 | 词语  总量 | 共用词语 | | 双区词语 | | 单区词语 | |
| 数量 | 比例 | 数量 | 比例 | 数量 | 比例 |
| 1 | >= 60% | 623 | 623 | 100.0 % | — | — | — | — |
| 2 | >= 65% | 873 | 872 | 99.9 % | 1 | .011 % | — | — |
| 3 | >= 70% | 1,222 | 1，220 | 99.8 % | 2 | .016 % | — | — |
| 4 | >= 75% | 1,737 | 1，732 | 99.7 % | 5 | .288 % | — | — |
| 5 | >= 80% | 2,548 | 2，537 | 99.6 % | 8 | .314 % | 3 | .118 % |
| 6 | >= 85% | 3,927 | 3，888 | 99.0 % | 29 | .738 % | 10 | .255 % |
| 7 | >= 90% | 6,564 | 6，341 | 96.3 % | 180 | 2.74 % | 43 | .655 % |
| 8 | >= 95% | 13,026 | 11，359 | 87.2 % | 1，236 | 9.49 % | 431 | 3.31 % |



以上统计显示出一个有说服力的证据：，在累计词频大于等于60 % 时，“共

用词语” 的走势和“词语总量”的走势几乎完全一样。从比例来说，在累计

词频（覆盖面）达到85 % 的3,927个词里，“共用词语”的比例一直高居99

%（这个数字已经囊括前表的“最常用词” 和“词常用词”的总和 — 3,039

个） ；而在覆盖面达到 95 % 的13,026 个词里，“共用词语”的比例还是

高居87 % （这个数字也已经远远超过前表的四类词语的总和 — 9,060个）。

它有力地显示出“共用词语”在汉语交际中是最主要最重要的部分。

4． 三点结论

综合以上统计和分析，我们大致可以认定以下三条结论：

1. 当代汉语的词语确实存在着地域差异。而“共用词语”无论在数量上（占

90 % 以上）， 使用频率上（在高频段和中频段），还是在累计频率上（覆

盖率达到 95 %） 都占了绝对的优势。

“双区词语” 和“单区词语” 数量上占了不到10 % 的比例，且大

都出现在低频段。其覆盖率处在 95 % 至 100 %　的区间里。所以汉语

词语的“内部一致性” 还是相当高的。京、台、港三个区域词语的相同

之处始终是主流。

1. 《汉库》中累计使用覆盖率到达95 % 的词语可以分为如下等级：

《汉库》各级词语的分级界标和实际数量

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 等 级 | 级 别 | 界 标 | 覆盖面 | 实际数量 |
| 最常用词 | 甲级词 | 1,000词次 | 70 % | 1,222个 |
| 次常用词 | 乙级词 | 4,000词词 | 85 % | 2,705个 |
| 常用词 | 丙级词 | 6,500词次 | 90 % | 2,637个 |
| 通用词 | 丁级词 | 13,000词次 | 95 % | 6,462个 |
| 共 计 |  |  |  | 13,026个 |

依照以上等级和级别的划分认定，各级词语分级的数量约为：甲级 1,222

个，乙级词 2,705 个，丙级词2,637 个，丁级词6,462 个。依其累计数量，

四级词语相互之间的累进比重大约为 1：4 ：7 ：13 。其中“最常用词” 、

“次常用词”和“常用词”属于“高频词” 和“中频词” ，而“通用词”

属于“低频词” 中的“高低频词” 。根据这个比例，可以认定，对京、港、

台三地的汉语词语分级定出下列分级标准是有事实根据，也是比较妥当的：

甲级词 1,200 个， 乙级词 2,500 个，

丙级词 2,500 个， 丁级词 6,500 个。

1. 《汉库》中“共用词语” 、“双区词语”和“单区词语”的数量和比例

在各词频段级的分布情况如下：

“共用词语” 、“双区词语”和“单区词语” 的数量和比例在各词频

段级的分布情况

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | 共用词语 | | 双区词语 | | 单区词语 | | 总 计 |
| 数量 | 比例 | 数量 | 比例 | 数量 | 比例 |
| 高频词 | 128 | 100.0 % | — | — | — | — | 128 |
| 中频词 | 6,216 | 93.6 % | 327 | 4.9 % | 96 | 1.5 % | 6,639 |
| 低低频 | 5,569 | 11.5 % | 10,474 | 21.7 % | 32,210 | 66.8 % | 48,253 |
| 总 计 | 16,742 | 27.5 % | 11,503 | 18.9 % | 32,566 | 53.5 % | 60,811 |

以上数据显示，“共用词语” 使用中的高频率和高覆盖率使得“华人文

化圈” 的汉语交流在百分之九十的程度上没有语言障碍。“共用词语” 依然

在九十年代的当代汉语中占了核心的地位和绝对的优势。其次，对说汉语的

中国人以及学习汉语的外国人来说，掌握了近 1,000 个“最常用词” 和

3,000 个“次常用词” ，是最低量的“词语集” 。而要使自己在京、台、

港三地工作、学习、生活、旅游时不遇到什么大的语言阻碍和问题的话，就

必须掌握前近 7,000个“常用词” 。但那也还只是中等量“词语集” 。

高等量的“词语集”应是 13,000 个左右。

“九十年代汉语词汇地域分布的定量研究”

[香港] 陈瑞端 汤志祥

A Quantitative Analysis on the Lexical Distribution in different Chinese Communities in 1990's

Chan shui-duen Tang Zhixiang

论文提要：

本研究以香港理工大学《中国大陆、台湾、香港汉语词库》的六百万字语料为基础，通过检索和统计，对九十年代通行于整个汉语·汉文化区域的汉语词汇进行定量分析研究。本文着重考察当代汉语词汇的“共同底层”和三个区域之间的“地域差异” ， 从数量特点及其分布角度，对流通于整个汉语区域的“基本词汇集”和仅在各自区域或者其中两个区域之间流通的“子词汇集”作一个初步的定性分析。

Abstract:

Based on the Chinese Corpus of The Hongkong Polytechnic University which was completed in June 1997 and contains over 5 million Chinese characters, the paper focuses on the common base of Modern Chinese words in 1990's, as well as their lexical varieties among different Chinese communities( eg.Mainland, Taiwan and Hongkong). The quantitative search by the concordance of the

Corpus and its corresponding analysis lead to an objective and qualitative conclusion on the lexical distribution in above-mentioned three regions.

联系地址：

汤志祥： 广东深圳大学文学院中文系 邮编：518060

香港理工大学中文及双语学系 QT 528室

1. 本论文在收集资料和撰写过程中得到香港理工大学语文教学中心（Chinese learning

   centre）主任陈瑞端教授（A. Professor Chan Shui-duen）的指导和帮助。特此鸣谢。 [↑](#footnote-ref-1)
2. “两岸”指“海峡两岸” ，即大陆和台湾。“三地”指大陆、台湾和港澳三个区域。 [↑](#footnote-ref-2)
3. 这是原香港理工大学中文及双语学系首席教授张日升 ( Chair - Professor Cheung Yat–

   shing ) 于一九九一年创办的。此处沿用原来的暂名。现任项目负责人为陈瑞端教授 。 [↑](#footnote-ref-3)
4. 该词库用“已切分单位”指称根据本身研究制定的切分条例的切分出来的语言单位。此处

   使用通行的术语“词条”和“词次” 。 [↑](#footnote-ref-4)
5. 有学者建议称之为“京港台汉语词库”或者“陆港台汉语词库” 。此处采用前者。 [↑](#footnote-ref-5)
6. 这里表示的是所收的一地区的语料量和三区域总的语料量之间的相对比例。后两项同。 [↑](#footnote-ref-6)
7. 比如辞书《大陆和台湾词语差异辞典》 邱质朴主编，南京大学出版社，1990年 [↑](#footnote-ref-7)
8. 比如辞书《香港话词典》 吴开斌著，花城出版社，1997年，广州。

   《香港话· 普通话词典》 朱永锴著，汉语大词典出版社，1997年，上海 [↑](#footnote-ref-8)
9. 比如辞书《当代港台用语词典》 朱广祁编著，上海辞书出版社，1994年 [↑](#footnote-ref-9)
10. 国内学术界存在着一个所谓“港台词语”的术语。这仅是一个以大陆为立足点进行观察的

    结果。因此实际上还应有“京台词语”和“京港词语”之说。况且，台湾词语和香港词语

    完全不能够混同一起说，其间既有各自的“区域词语”之分，也有闽、粤方言之分。 [↑](#footnote-ref-10)
11. 根据《中港台汉语语料库》的切分条例，凡是音节长度超过四个音节的语言单位都要

    硬性进行切分，所以《中港台汉语语料库》并没有五音节或以上词长的词语。下同。 [↑](#footnote-ref-11)
12. 《汉库》里外语字母看成为一个音节的单位。故“卡拉OK”算作“四音节词语”。 [↑](#footnote-ref-12)
13. 现在例举的是九十年代前期的情况。不包括在过往的时间里因三个区域的人员的交往而引

    致的词语互相交融和吸收的结果。 [↑](#footnote-ref-13)
14. 因为粤港两地均通行粤语，而台湾不通行，因此下表中包含有一部分粤语词语。 [↑](#footnote-ref-14)
15. 所谓“人工干预” ，也叫做“专家干预” 。它指的是：请专家用专业“定性”的办法，

    根据经验和辅助材料对语料进行人为的人工方法判别，试图纠正因为纯电脑统计中语料

    数量不足或语料偏差所造成的“误差” 。（参见《汉语水平词汇与汉字等级大纲》13页

    至15页，国家对外汉语教学领导小组办公室汉语水平考试部编，北京语言学院出版社，

    1992年6月）。 [↑](#footnote-ref-15)
16. 北京航空航天大学汉语预料库库量近两千一百五十万。台北中央研究院资讯科学科学研究

    所中央研究院历史语言研究所新闻语料库库量近一千五百万。而《中港台汉语词库》总库

    量为五百万。其中每一个区域的分语料库量大约是一百三十多万。这个数量现在看来显然

    无论从相对量和绝对量来说都是很不足够的。 [↑](#footnote-ref-16)
17. 将《中港台汉语词库》的高频、中频、低频的分界分别划在3，000 次、300 次和30 次之

    处，是因为这个词库和其他的单区语词库相比，实际上是一个有三个独立的单区语分词库

    相加的总词库。按照通行的频率划分法，单区语词库的高频词的频率为 1，000 次，低频

    词的频率为 5 次。因此，上述画法仅是将这个三区总词库的数值相应扩大了三倍，以便

    划分标准彼此一致，方便两者之间进行比较、分析。而把每一级再细分为二，是为了将统

    计做的得更细致、更详尽。 [↑](#footnote-ref-17)
18. 北京语言学院出版社，1992年6月. [↑](#footnote-ref-18)
19. 这里是指经过专家“人工干预”后的实际数量。 [↑](#footnote-ref-19)
20. 此比例引自《现代汉语频率辞典》北京语言学院出版社1986年 ，原文并无。 [↑](#footnote-ref-20)