《现代蒙古语词频统计系统》的设计与实现

吉日木图、嘎日迪、赛音、白晓玲
（内蒙古自治区电子计算中心）
达·巴特尔
（内蒙古自治区社会科学院）
阿拉腾敖道
（内蒙古自治区蒙古语文工作委员会）

摘　要： 本文介绍了《现代蒙古语词频统计系统》的设计原理与实现方法。重点论述了统计的目的和意义、统计语料的选取、统计模型的建立、B—树数据库的实现、同形词标记、单词独形统计、单词变形归一统计、复合词自动切分统计、结果合并、频度排序、使用度排序、读音排序以及结果输出等语料处理和统计软件系统的设计与实现过程中遇到的问题和采用的技术。

关键词： 现代蒙古语　词频　统计　B-树　数据库　排序

Design and Realization of the Statistical Analysing System of Modern Mongolian Word Frequency

Jirumtu, Gardi, Saiyin, Bai xiaoling
(Inner Mongolian Autonomous Regional Electronic Computing Center)
Da · Baatar
(Inner Mongolian Autonomous Regional Academy of Social Sciences)
Altanodo
(Mongolian Language Working Commission of Inner Mongolian Autonomous Region)

Abstract: This article introduced the designing principle and realizing method of the statistical analysing software system of modern mongolian word frequency, especially, introduced the techniques and method of solving several problems on mongolian word frequency calculation as well as statistical analysis such as creating statistical model, designing B-tree database system, marking homographic word, monographic and compund word counting and merging, sorting results by word frequency, word usage and word pronunciation.

Key words: Modern Mongolian, Word Frequency, Statistics, B-tree, Database, Sorting

一、目的和意义

蒙古民族拥有辉煌的历史、灿烂的文化、丰富的语言、独特的文字。近千年的文字史为我们留下了很多文献古籍。专家估计蒙古语词汇约有10万左右。这么庞大词汇量，任何人也不可能完全掌

握，而且也不是每个人都有这个必要。因此很有必要从词汇应用的差异方面把蒙古语词和词汇进行一次等级划分，按词和词汇的使用度加以区别。这样，人们就能有效地用其所用，不耗费很多的时间和精力。也就是解决存储冗余问题，致使提高信息传递的质量。这是《现代蒙古语词频统计》的第一个目标和意义。其次，从识字教学的角度讲，首当其冲地给儿童们、小学生们解决最基本的、最常用的词汇问题。另外，信息时代要求语言研究沿着精密的方向发展。从理论和实践上都对蒙古语词汇计量研究提出了词频统计的要求。特别是信息处理、情报检索、机器翻译、人工智能、文字改革以及语言学、教育学、心理学研究等都需要经过科学统计和客观筛选的蒙古语词频统计数据为重要参考。

我们针对社会生活、语言教学、信息科学等方面对蒙古语词频统计的需要，运用对各种题材、体裁的蒙古语材料进行抽样统计的方法，对现代蒙古语的词汇的实际使用情况做了较全面的调查研究。根据词在各类语料中的出现频率和使用度比较客观地展示出蒙古语词汇的使用概貌，并为他们划分出常用等级。

二、工程设计与语料抽样

大家知道，对一种语言不确定范围或不加限制地对所有材料进行统计是不可能的。因此《现代蒙古语词频统计》工程中，采用抽样统计的手段，使其在类型、范围、质量、数量方面既满足普遍性，又强调规范性对语料进行了选择。

首先，要在时间和空间上对所选语料限定范围。蒙古语是一个跨国界的国际性语言（中国、蒙古国、俄罗斯），并且已有近千年的文字史。所以，我们把统计语料的起始时间确定为1949年建国（也包括自治区成立初期）以后，把收集语料的地域范围限定为中国境内。这样的限定主要是为将来可能进行的蒙古国和中国、俄罗斯和中国境内蒙古语词汇对比研究留有余地。

其次，在质量方面要求所选语料符合中小学或成人以及对外蒙古语教学对一般常用词的需要。也就是强调普遍性，而不偏重专业性；注意规范性，而不照顾特殊性。所以选择语料时面要广，点要精。同时还特别注意到了著名作家的优秀作品（获奖作品）和规范的教科书。与此相反，对于那些乌七八糟的劣质语料坚决杜之门外。

再次，在数量方面要求所选语料既可靠又经济，也就是说在合理性和准确性之间找到一个最佳限量。这方面我们在参考了国内外筛选各种自然语言常用词经验的基础上把限量数目确定为不少于百万词，（包括独形附加成分在内）不超过二百万词的语料。

最后，在类型方面要求所选语料要有一定的代表性和客观性。即能够客观地反映口语，科普·政论，文学四大语体特征和代表各类创作人员的用词特点。如，教师、编辑、翻译、公务员、作家、科普作家、女作家等等。以此来保证其统计结果的普遍意义和覆盖面。

另外，我们在课题酝酿过程中意识到了社会现状、文化传统、文献古籍、社会层次以及文化、科技、经济的发展水平等对一个民族语言的词汇范围和现状所起的作用。并综合上述把我国境内《现代蒙古语词频统计》的语料按照口语10%、科普20%、政论30%、文学40%的比例分为四大类输入微机进行处理。我们的这种比例抽样考虑了我国境内蒙古语和我国现行蒙古文的以下几个特点：第一，有一定数量的文献古籍；第二，翻译语言占一定比例；第三，口语和书面语差异大；第四，地区性通用语。但是，有必要说明的是：由于中小学蒙古《语文》二十二册（除诗词、古文、蒙古国作家撰文外）分别插入四大类之后原有的1:2:3:4比例有所变动，即口语1/13、科普2/13、政论3/13、文学

7/13。我们的统计结果是在变动后的语科基础上产生的。

三、统计中的若干问题及技术技巧

蒙古文同其它文字有明显的区别，即从形式上看蒙古文属于连写式拼音文字。词与词之间用空格分开。因此对蒙古文词的独形的频次(形次)统计来说是很容易的。但从语言学的角度考虑独形中包含了很多如附加成份以及词的变形等非词的部分。也有多个独形组合在一起，构成一个词汇的问题。因此解决这些问题大大增加了统计的难度。

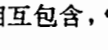
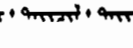
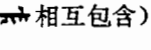
1、同形词处理

《现代蒙古语词频统计》首先要解决的问题是蒙古文同形词问题。同形词的处理上我们采用了人机对话形式(机器自动定位、人工区分标记)的预先标记的方法。原因是蒙古文同形词问题靠计算机自动处理目前还不够现实。蒙古文中的50%左右字母存在二义性，用机器是无法辩认。这是我国现行蒙古文本身存在的一大弊病。对同形词处理的具体作法是：

建立开放性同形词词典(即随时增加词条的计算机词典)。

操作员用同形词标记软件对语料中的同形词根据语境区分词义进行标记。标记的思想是同形词标记软件从文件的头开始逐条词在同形词典中查找匹配，确定是否是同形词，若认为是同形词则屏幕上提示同形词的不同词义进行选择区分标记。

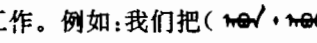
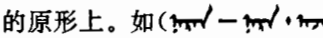
词频统计时对已标记的同形词认为是不同的词，分别对待统计和计算。

在实际操作中发现有些同形词是动词。蒙古语动词具有很多变形，不可能每个变形都加到词典中，那样词典太庞大了(可能出现的词尾变化就有4多万条)。因此对同形动词不仅要匹配词干，还要匹配词尾。更有甚者，有些同形动词包含其它同形动词一个套一个，增加了匹配的难度和准确性。为此采取了同形动词的三级词典模式。第三级同形词动词中同形词词干部分包含了第二级和第一级同形动词库中个别同形动的词干，同样第二级同形词动词的词干部分包含了第一级的个别同形动词的词干。在同形词定位标记的过程中程序从第三级同形动词词典开始依次第二级、第一级同形动词词典和普通同形词词典中以最大匹配原则正确定位同形词，并进行标记区分。(如：  相互包含，  相互包含，  相互包含)

2、人名地名处理

在语料中对人名地名认为是非词而不统计。为了区分人名地名，预先标记处理。对每一个人名地名用一个特殊的专用符号括起来，统计时跳过标记的人名和地名。

3、词尾变形处理

蒙古语是主要由追加在词根或词干上的附加成分来表示词的派生意义和语法意义。特别是动词的体、态、式、时以及人称和名词代词的词尾变化也很活跃。为了最终以词为统计单位来体现，还原蒙古语词尾变形是必需要做的工作。例如：我们把()等都还原到()上，其统计结果是合并计算的每一个变形形次的累计。另外，名词的变形和复数、代词的变形等也都还原到该词的原形上。如()等。

对词尾变形的处理上采用的方法是首先建立一个词尾变形还原对照词典库，分别对每一条语料中出现的变形词做对照转换表。其次，对原语料独形统计得到该文件的独形频次表。最后利用变形还原对照词典库对独形频次表中的变形词进行频次数合并到原形词的频次数上(同时删除该变形)得到每个语料文件的单词频次表。

4、复合词处理

对于蒙古文材料而言从形式上根本无法区分单词和复合词。若事先不做工作是很难进行复合词的确定和词频的有效统计。特别是蒙古语的复合词的形式、划分方法、规范以及整理等诸多理论问题尚未明确的今天，其难度是可想而知的。为此首先要确定复合词统计的范围，我们采取了一个即有可操作性又能被大家普遍接受的方法，即把已出现的蒙古语辞书所收列的复合词汇集为一体建立了一个复合词辞典。

确定了复合词的范围之后提出了如何统计复合词的问题。我们课题组人力、物力都比较弱，不可能人工切分，而只能采用词典对照机器自动切分的方法。在微机上做具有叁万多词条的词典并对照切分复合词，遇到了容量与速度的矛盾问题，经过精心组织和设计完成了比较有效的复合词统计软件。

另外，用复合词词典对照方式自动统计复合词时，也遇到了复合词的词尾变形问题。对此也特做了近四万多个可能出现的动词词尾形式库。可通过对复合动词词尾匹配来确定当前统计的词是不是该复合动词的变形。

复合词自动切分统计完之后进行复合词合并。复合词的合并是对已经统计得到的单词统计文件和复合词的统计文件结果进行合并得到准确的词汇统计结果文件。合并过程中由于独形统计、变形合并的过程中未考虑复合词的问题，而把复合词的每个独立的形式都按一个单词来统计，而复合词统计中只考虑了匹配到的复合词的频次而未考虑单词的频次，因此，复合词合并时，对每个构成复合词的单词逐词查找其统计结果，并从其频次中减去该复合词的统计频次。最后该复合词及其统计结果追加到单词统计结果中，得到词汇统计文件。

5、读音排序

我国现行蒙古文(回纥蒙古文)的字母中有 14 个字母的 19 个字形在词的不同位置带有二义性和歧义性。致使采用形码体系的系统不可能实现直接用计算机自动读音排序，这是蒙古文本身的弊病所造成的。为了完成读音排序(常规词典排序)的任务采用建立读音对照词典的方法。首先根据统计得到的总词表建立读音对照词典。在建立过程中由程序自动给出每一条词的读音，若读音有误由人工修改即可。建立好读音对照库之后，可以对统计结果的每一条词汇都按其读音进行排序得到读音排序词汇词典。

6、频度排序

频度排序是指按每条词汇的使用频次的大小进行的排序，这一排序是只考虑其出现频率而不考虑其分布情况的排序。同一频度(词次)级内按读音顺序排序。

7、使用度排序

使用度的概念是综合了频次、篇章数和类数三方面的因素，计算得到的词次，从数值上可以看出词条在语料中的使用程度和散布情况。频次与使用度越接近说明该词分布越均匀。同一使用度级内按读音顺序排序。

使用度的计算公式如下：

$$Uk = DEk \times Tk$$

$$DEk = \begin{cases} \frac{1}{2}Dk + \frac{1}{2}DIk, & Fk \geqslant 0.0001 \text{ 时} \\ DIk, & Fk < 0.0001 \text{ 时} \end{cases}$$

$$Dk = 1 - Sk/(Fk \times (n-1)^{\frac{1}{2}});$$

$$Sk = (\sum_{i=1}^{n}(Fk_i - Fk)^2)^{\frac{1}{2}}/n$$

$$Dik=(Pk+80Lk+480)/1122$$

其中：Uk 是使用度；

Tk 代表 k 号词的统计词次；

Fki 表示 k 号词在第 i 类语体里的相对频率；

Fk 表示 k 号词在各类中的平均频率；

n 为语体类数；

Pk 表示 k 号词的分布篇数；

Lk 表示 k 号词的分布类数；

Dk 及 DIk 均称为 k 号词的散布系数；

DEk 是总散布系数；

8、分布排序

分布排序是按词或词汇出现的类数、篇章和词次进行的排序。先按类数大小排序，在同一类数级内按篇章数大小排序，在同样类数篇章数的情况下按词次大小排序，若类数、篇数、词次都一样则按读音顺序排序。根据此结果得到现代蒙古文的分布最广的词语。

9、编差分析

对每条词的频次和其分布情况进行分析计算得到其实际分布与理想分布之间的偏差。

偏差的计算公式如下：

$$Ek=(\sum_{i=1}^{4}(ok_i-ek_i)^2/ek \qquad ek_i=ek\times P_i$$

其中：ek 表示第 k 号词在综合统计中的出现次数，P_i 表示第 i 类语料在综合语料中所占的比例，ek_i 称为 k 号词的在第 i 类中的期望值词次，ok_i 是 k 号词在第 i 类中实际出现的统计词次，Ek 为第 k 号词的偏差系数。

四、统计模型与统计步骤

1、统计模型

《现代蒙语词频统计》工程是一个比较复杂的工程，对约 100 万词的语料中进行统计每条词汇的出现频次以及出现的篇章数。最后计算各词汇的频度、使用度，分析其分布情况。因此具有量大、过程繁杂等特点。为了能够保质、保量、按时完成这一项任务，在设计统计模型方面下了很大功夫，经统计、实验、改进形成了如下的分层统计模型。

在第一层对语料文件进行同形词标记，人名地名标记等预处理。

在第二层对已预处理好的语料文件分别进行单词统计、复合词统计和单词变形合并、复合词合并，形成对应于每篇文章的词汇频次表。

第三层对词汇频度表按分类进行合并得到分类词汇频度表。

第四层对分类统计结果合并得到词汇统计结果。

最后一层对所得结果进行排序得到频度排序词表、使用度排序词表、读音排序词表等。

《现代蒙语词频统计》软件采用工程化管理方法，每进行一步都对已处理文件进行记录，以便检查漏统计和避免重复统计。统计过程以批量形式对每个文件分别进行处理。

2、数据库实现

《现代蒙语词频统计》的数据库采用的是 B—树(B-tree)数据库。B—树数据库与关系数据库完

全不同，数据库的结构同关系数据库的 B—树索引文件一样，集数据库与索引为一体的数据库。

B—树数据库有：a)数据记录可以不定长度，随记录内容的大小，可以任意变化，适合于词汇的长度不一的情况。b)定位速度快，查找方便。c)数据库与索引合到一起，能够节省空间等优点。

为了进一步加快搜索与处理速度把 B—树数据库设计为可调入扩展内存中变成内存数据库。在具有 4M 内存的机器上大部分数据库都可以调入扩展内存运行。但为了达到自适应机器配置，数据库在打开时自动检测扩展内存大小并以其尺寸大小限制调入扩展内存。其余部分仍在硬盘上存放和使用。数据库关闭时自动把扩展内存内容存入硬盘上文件。

3、统计步骤

a)语料预处理——首先把语料作品录入计算机，并进行校对无误后进行语料预处理，预处理包括：人名地名的标记、页码插入、同形词标记。标记方法前面已详述，在此不再赘述。

b)独形频次统计——对每个语料文件进行独形频次统计得到独形统计文件。

c)变形对照转换——对统计得到的独形统计文件进行变形对照转换得到单词统计文件。

d)复合词统计——对每个语料文件用复合词匹配方法进行复合词统计得到复合词统计文件。

e)复合词合并——复合词合并是对已得到的单词统计文件和复合词统计文件进行合并，得到词汇统计文件。

f)词汇分类合并——对已统计合并得到词汇统计文件，按不同类型(即口语、科普、政论、文学四大类)进行合并。合并过程中不仅把每条词的词次累积得到总词次，而且还要对每篇文章中出现的词的篇章数累加递增的方法得到该词在此类中的分布篇章数。如此得到分类统计结果词典。

g)词汇统计结果合并——对已得到的分类统计结果词典进一步合并得到词汇统计结果词典。

h)结果排序——对词汇统计结果词典进行读音排序、频度排序、使用度排序、分布排序、偏差分析分别得到词汇读音排序、词典词汇频度排序词典、词汇使用度排序词典、词汇分布排序词典、词汇偏差分析词典。对词汇分类统计结果词典进行按频度排序得到分类频度排序词典。

i)结果输出——对统计结果的打印输出采用直接形成华光排版小样文件输出的方法。这样输出的结果直接由华光排版系统排印输出。

同样对独形统计文件(步骤 b 的结果)做步骤 f、g、h、i 的操作，可得到独形的各种统计排序结果。

五、结束语

在《现代蒙语词频统计系统》的设计和开发过程中我们始终遵循快速、准确、方便的宗旨。最大限度地利用机器资源，减少人工干预。在统计和合并过程都采用了工程化管理，对已处理文件进行登记，自动检测文件的重复统计和合并。统计合并全部都是批量化，在处理过程中无需人工干预。在需要人工处理的地方尽量提供方便的操作界面和帮助措施达到运用灵活自如的目的。

参考文献

1、北京语言学院语言教学研究所编，《现代汉语频率词典》，1985 年，北京语言学院出版社。

2、陈原主编，《现代汉语定量分析》，1989 年，上海教育出版社。

3、Willian James Hunt，《The C Toolbox》，1985.