◆ 国家社会科学基金项目：
《藏文词汇频度统计及其应用研究》(98BYY020)
◆ 清华大学智能技术与系统国家重点实验室基金项目：
《大型藏文语料库建设及其应用研究》（0105）
◆ 国家社会科学基金重点项目：
《藏语语料库建设研究》(05AYY001)
◆ “西北民族大学学术成果出版基金”资助出版

ཡིག་དཔྱད་ཀྱི་ཚད་ལྡན་ཆེན་མོ།།

现代藏文频率词典

主　编◎卢亚军
副主编◎高　瑞　马少平　马进武
编　委◎张　敏　夏力耕　卢靖华
罗　广　扎西次仁　达尔吉　贺　胜

民族出版社
མི་རིགས་དཔེ་སྐྲུན་ཁང་།

◆ 国家社会科学基金项目:
《藏文词汇频度统计及其应用研究》(98BYY020)
◆ 清华大学智能技术与系统国家重点实验室基金项目:
《大型藏文语料库建设及其应用研究》(0105)
◆ 国家社会科学基金重点项目:
《藏语语料库建设研究》(05AYY001)
◆ “西北民族大学学术成果出版基金”资助出版

༄༅།།དེད་དབས་བོད་ཡིག་འགྲུལ་ཚད་ཀྱི་མིང་ཚིག།

现代藏文频率词典

主　编◎卢亚军
副主编◎高　瑞　马少平　马进武
编　委◎张　敏　夏力耕　卢婧华
罗　广　扎西次仁　达尔吉　贺　胜

民族出版社
མི་རིགས་དཔེ་སྐྲུན་ཁང་།

图书在版编目(CIP)数据

现代藏文频率词典 /卢亚军主编. －北京:民族出版社,2007.7
ISBN 978－7－105－08440－1
Ⅰ.现… Ⅱ.卢… Ⅲ.藏语－词典 Ⅳ.H214.6

中国版本图书馆 CIP 数据核字(2007)第 112856 号

民族出版社出版发行
(北京市和平里北街 14 号 邮编 100013)
http: /www.mzcbs.com
北京迪鑫印刷厂印刷
各地新华书店经销
(藏编室电话:010－64275311 藏文发行科电话:010－64227665)
2007 年 10 月第 1 版 2007 年 10 月北京第 1 次印刷
开本:880 毫米×1230 毫米 1/16 印张:87.50 字数:1577 千字
印数:0001－300 册
ISBN 978－7－105－08440－1 /H·617(藏 66)
定价:260.00 元

ཚོགས་སྤྱིག་འགན་འཛིན་བ། འབྲི་མོ་སྐྱིད། སྒྲོལ་མ་སྐྱིད།
责 任 编 辑： 周毛吉 卓玛吉

㊣ ། ། དེད་རབས་བོད་ཡིག་བསྒྲུར་ཚད་ཀྱི་མིང་མཚོད། །

ཚོམ་པ་ཅི།	ཀློ་ཡངས་རྒྱལ།
དཔེ་སྐྲུན་མཁན།	མི་རིགས་དཔེ་སྐྲུན་ཁང་།
འཕྲིམས་སྒེལ་མཁན།	མི་རིགས་དཔེ་སྐྲུན་ཁང་གི་བོད་ཡིག་ཚོམ་སྤྱི་ཁང་།
གནས་ཡུལ།	པེ་ཅིན་ཇོ་ཕིང་ལི་བྱང་ལམ་ཨང་14པ།
སྒྲུག་ཁྱིད་ཨང་གྲངས།	100013
ཁ་པར་ཨང་གྲངས།	010－64275311
བརྒྱད་འཚོང་མཁན།	རྒྱལ་ཡོངས་ས་གནས་སོ་སོའི་ཞེན་ཀུ་དཔེ་ཚོང་ཁང་།
པར་འདེབས་མཁན།	པེ་ཅིན་ཏེ་ཞེན་པར་ཁང་ནས་པར་བཏབ།
དེབ་ཚད།	880mm×1230mm 1 /16
པར་ཤོག	87.50
པར་གཞིའི་སྤྱི་ཟླངས།	2007ལོའི་ཟླ་10པར་པར་གཞི་དང་པོ་བསྒྲུགས།
པར་ཟློངས།	2007ལོའི་ཟླ་10པར་པེ་ཅིན་དུ་པར་ཟློངས་དང་པོ་བརྒྱབ།
པར་གྲངས།	0001－300

ISBN 978－7－105－08440－1 /H·617(藏 66)
དེན་གོང་། སྒོར་260.00

代　序

——基于大型藏文语料库的藏文字符、部件、音节、词汇频度与通用度统计及其应用研究

[摘　要] 本文在考察、借鉴上百年来国内外对德、英、汉、蒙、藏等语种所作的文字计量研究，特别是各种汉文字词频度统计研究成果的基础上，论述了基于大型藏文语料库的藏文字符、部件、音节、词汇频度与通用度统计、分析的理论与方法及其实现过程．该项研究是一项基础性研究，首次实现了大规模藏文词汇频度与通用度统计，具有一定的创新性．其成果有助于藏族基础教育和扫盲教育的语言文字教学，对藏语语言学研究和藏文信息处理领域的许多方面具有重要应用价值．

藏族有着悠久的历史、灿烂的文化，其卷帙浩繁的文献遗存在国内仅次于汉族，在世界民族之林亦非同凡响．藏文约自公元7世纪松赞干布时期创制以来，在其日臻完善的1300多年的历程中，倾注了历代诸多先哲和学者的聪明智慧．在信息化时代到来的今天，在党和国家正确的民族政策和民族语言文字政策下，借助先进的计算机技术和语言学理论的新成就及其研究方法，对藏文进行量化统计、分析与应用研究，一方面对提高藏族中小学教学水平和民族教育质量有着显而易见的效果和作用；另一方面对藏文信息处理技术的发展也具有重要的作用和意义．

本课题研究在遵循并应用计算语言学、语料库语言学、分词规范与方法、藏文文法、数理统计方法和一些有关的最新理论研究成果和方法的基础上，以数量级足够大的“七分一总”语料库为统计样本，以从《藏汉大辞典》等8部词典和正字法中选取的词汇为机载统计用《词表》，以计算机和人工相结合的方式，以藏文字符、部件、音节、词汇及其频度、频率、累计频率和词汇通用度为统计、分析与研究对象，力求为藏文字符、部件、音节和词汇的语言学研究、教学研究以及信息处理领域相关的应用研究，提供干扰因素尽可能少，信息量尽可能大，客观性尽可能强，适用性和可靠性尽可能高的统计分析数据依据和第一手资料．

本课题基于大型藏文语料库的藏文字符、部件、音节和词汇等这几个层次的频度统计与分析，其研究成果的应用领域涉及以下几个方面：

(1) 字符　用于信息交换用藏文字符编码、文字识别研究及软件开发．

(2) 部件　从字符统计结果中析出藏文部件，为藏文的属性研究、信息交换用藏文部件编码以及藏文键盘布局与输入法研究提供基本参数．

(3) 音节　用于藏文语素、构词法、语音识别、自动校对等研究及软件开发．

(4) 词汇　主要用于《现代藏文频率词典》《藏文常用词汇词典》(藏汉双解)《藏文中小学生常用词表》《扫盲词表》及配套用书等词典、词表的编纂，以及藏文词汇级次研究、藏文整词输入法、藏文自动侦错校对、文字识别、语音识别、情报检索、机器翻译、词汇学、语言学、认知学研究等．

1　藏文语料库的语料采集、分类、抽样和预处理

1.1　语料采集

词汇是一种语言里所有的(或特定范围的)词和固定短语的总和[1]. 藏文词汇频度统计、研究的对象,首先应当是现实生活中人们所能经常接触到的语言文字材料;其次,由于基本词汇具有稳固性、能产性和全民常用性,所以一定数量不同种类的历代古典文献语料也是不可或缺的.

本研究项目的语料采集采取"厚今薄古"与"古为今用"互补的原则."厚今"是为了体现词汇的"全民常用性"和"能产性";"鉴古"是为了体现词汇特别是语素的"稳固性". 由于藏族社会在现代科学技术领域相对落后,而在传统文化领域(即所谓"大小五明"或称"十明"),特别是佛学、史学方面又相当发达,所以"古为今用"是十分必要的.

按照本课题研究所制定的语料采集的原则和方案,我们从全国各出版、印刷机构的藏文图书征订单和内容简介中,按反映现实生活、传统文化和自然科学三个大的方面进行选目、征集. 在具体的语料采集过程中,得到了有关方面人士及部分作者的大力支持. 经过3年多的努力,我们采集到的各种语料总数达1.3亿字节,相当于4300多万音节\字.其中有日积月累的报刊,鸿篇巨制的文集,分门别类的著述,大中专文理科教材等等,可谓五花八门,丰富多彩,而且校对也不存在问题. 语料的数量级、代表性、客观性和覆盖面完全可以达到词频统计在理论和实践上的要求(语料篇目详见本课题研究报告附录).

1.2 语料分类

国内外有关词频统计方面的研究,对于不同文种在不同的历史时期和条件下,根据不同的研究方法、手段和目的,其语料分类各有千秋,不一而足,既有共性,也有个性. 本课题组在考察并参照许多种语料分类的基础上,在全面分析藏族文化的历史和现状的前提下,根据本课题研究的任务和目的,将语料具体划分为"七分一总".所谓"七分"即报刊类、文学类、教育类、科技类、佛学类、历史类和传统文化五明类等7个分类语料库;"一总"为7个分类语料库"七合一"的汇总语料库.

1.3 语料整理

语料的整理分以下几个步骤:

(1) 语料分类　依据本课题研究所制定的分类方法,在对语料进行领域分类的基础上,再按作品的文体、风格、年代、篇幅进行人工遴选. 这也是一个抽样过程,即把缺乏代表性的语料,如《××图书目录》《雪域历代名人词典》《红教密咒》等作品人为地抽取掉,而把认定为适合入选语料库的电子文本,逐页逐行、逐字逐句进行审查,并对文本按篇目进行标记,然后将原始语料进行归类,使之成为原始的分类生语料库.

(2) 文本预处理　对7个分类语料库的原始文本,需要按藏文频度统计的要求进行文本预处理. 具体的处理内容和方法如下:

① 合并文本文件　对所有文本文件进行链接式合并,即把3000多篇报刊文本合并成一个报刊分类文件;分别按分类把几十部图书文本合并成该分类的一个文件.

② 藏文隔字符与隔句符(详见下文).

③ 排版控制符和空格　电子文本中的排版语言控制符在打印、发排文稿中不打印出,但可屏显并占文件字节数. 所以,所有排版控制符包括软、硬换行符和软、硬空格一概予以人工干预删除. 只有统计样本中的字符总数受干扰越小,字符和部件的频次及频率的统计准确率才能越高.

1.4 语料抽样

样本分为两种类型:一类是7个分类语料库各为独立的样本集合;另一类是7个分类语料库合并的汇总语料库为一个样本集合.

样本抽样采取了两种形式:一种是按分类对作品的文体、风格、年代、篇幅进行篇目遴选时,把缺乏代表性的语料人为地抽取掉;另一种是对各类经过预处理的语料进行随机等距抽样,即从各类

经过预处理的约 1 200 万～1 600 万字节的语料中，均等抽取 1 000 万字节的语料作为分类统计用样本．7 个分类统计用样本的合并文件即为 7 000 万字节的汇总统计用样本．

2 藏文词汇频度统计用《词表》的制定

“词表”也叫“底表”，具有机载电子词典的功能．主要用于对语料库的词条匹配、检索与词汇频度统计．以往对藏文所作的有关统计方面的研究，仅限于藏文字符和音节，而未见对词汇有统计研究．由于藏文词汇频度统计，一要建立一个规模足够大，至少也要达到七八百万音节（约两千多万字节）的电子文本生语料库；二要对语料库进行分词处理，这项工作迄今尚未见可资使用的成果；三要有一个词频统计用《词表》，而词表的制定又并非易事．本课题研究词频统计用《词表》的制定以及选词标准分述如下：

2.1 《词表》制定的步骤

（1）以《藏汉大辞典》（上、中、下）为本《词表》选词之蓝本．该辞典是一部大型的“以语词为主兼收百科的综合性藏汉双解辞书，全书所收词目共约 53 000 余条”[2]．该辞典的所有词条，由著名藏族语言学家西北民族大学马进武教授按选词标准确定入选词条，并对其中的异读词和错别词作了相应的处理，随后对入选词条用北大方正藏文字处理系统进行人工录入和反复校对．

（2）《词表》的补遗　尽管《藏汉大辞典》是一部权威的大型辞典，但其中也难免存在有所疏漏的问题．为了最大限度地提高《词表》的词汇覆盖率和可靠性，我们又选用《格西曲札藏文辞典》等 8 种[3]各具特色的藏文词典和正字法，分别对本《词表》逐词逐条地进行对照排查，拾遗补阙．这一步骤的工作十分细致而繁杂，但很有必要，也很有价值．本《词表》共录入 46 000 余词条，增补基本词汇和常用词汇 310 余条，最终定稿为 34 141 条．

（3）《词表》的校对　对《词表》的校对我们采取了三种方式．一是常规校对，即先过 1 至 4 校；二是在用前述 8 种词典和正字法分别对照《词表》增补词条的同时进行 5 校；三是组织三组校对人员❷分别对《词表》进行了第 6 校，等于说又过了 3 遍；最后一校第 7 校由课题负责人完成并审定．

2.2 《词表》的选词标准

本《词表》主要用于藏文词汇频度统计．根据《词表》的用途我们对选词标准作了如下规定：

（1）选词来源　以《藏汉大辞典》为主，辅之以《格西曲札》《新编藏文字典》等 7 种词典，另外还参考了一些数学、物理、化学、生物等自然科学方面的专业词典[4]和《汉藏词集》新词术语单行本（青海省政府编译室编印）等．

（2）理论依据　参考了近年来语料库学语言学、词汇学和字词频度统计研究方面的论文、专著．词和短语的选词原则与方法参照了《信息处理用现代汉语分词规范及自动分词方法》[5]《现代汉语语法信息词典详解》[6]《中文文本自动分词和标注》[7]和 GB10112－88《确立术语的一般原则与方法》[8]，以及《藏语计算机自动分词的基本规则》[9]等专著、国家标准和论文．

2.3 选词范围

1）全选词条　《藏汉大辞典》以及其它 7 种词典中的备选词条，无论属于藏族传统学科分类之“十明”中的哪一类，无论是古词还是今词，无论是书面语、口语、敬语、藻饰语还是方言，凡符合下列条件的就全部入选．

① 单纯词：单音节、双音节和多音节的单纯词全选，其中包括音译词和梵音词．藏文的单音节单纯词大多是合成词的词根或语素．单纯词的统计可为全面了解藏文语素的状况，尤其是其中使用频度高、构词能力强的语素提供研究数据．

② 合成词：双音节和三音节的合成词均入选．这类合成词是藏语词汇系统的主体，也是词频

统计的主要对象,其中的高频词是本课题研究的重点.

③ 其他:凡属划分成语素或合成词后,会失去原有词语意义的多音节词汇,像成语、熟语、缩略词、合称词等均入选.

2)未选词条　凡人名、地名(包括山名、水名、寺名)、书名和一些在结构上可划分为词或短语的专名、术语一般不选,但其中一些比较常见的可适当入选.

3)短语词条　对短语或词组(包括术语)词条的入选作了比较严格的限制,具体规定有以下三条:

① 短语词条的选取　短语是词和词按照一定语法关系结合起来的语言单位.本《词表》收入短语一是为了词频统计的需要,二是为了自动分词的需要.一般来说,若短语划分成语素或合成词后不影响对语义理解的不予收入,如"计划经计";而由语法虚词关联的短语,若是划分成词后会失去原有意义的则予以收入,如"出乎意料".

② 凡属"十明"范畴的术语,无论是词还是短语一般都选取,而过于冗长的则不选.

③ 属自然科学和社会科学领域的专业术语和新词术语,凡《藏汉大辞典》上有的均选取.另外,从数学、物理、化学、生物等专业词典中也遴选增补了一部分.

2.4　对《词表》的相关处理

1) 给定词条匹配条件　给《词表》中的每一个词条都给出一个匹配限定条件,即在每一词条的前后都加上隔字符(' ~'),这样便可保证该词条在电子文本中与相同词条惟一性的准确匹配

2) 词的后缀处理　藏语的大部分动词和少量的形容词、名词、数词的原形都带有后缀(ཨིང་ཟཁའ།)ཡ、ཏ、ཡི、ཡི、ས、སི等,但在真实文本中像ཡ、ཏ一般都很少带.另外,在西藏方言区,后缀ཡ、ཏ一般没有严格按照语法(特指音势论〈ཀྲག་ཟའི་འཇུག〉)添加.因此,《词表》中所有词条原形的后缀ཡ、ཏ都要去掉,而像为数不多的带有ཡི、ཡི、ས、སི的词条,也专门设定了带后缀和不带后缀的两种,以便这些词条在真实文本中能准确匹配.

3)粘着型词缀的处理　粘着型词缀在藏文文本中大量存在,而在任何一部词典中都没有其词条.粘着型词缀大多是动词、名词、形容词的后缀与某些语法虚词紧密结合的词型,它只有语法意义而没有词汇意义,不能单用.为了把藏语中粘着型词缀现象也能正确地统计并反映出来,我们在《词表》中增加了ངས、ཀོས、དེས、ཡས、བས(ཅིད་ལྡན)和ཝའི、བའི、ཡའི、ཏའི、མའི、སའི、ངའི、ཀའི、དེའི(འཇིལ་ལྡན)等粘着型词缀词条.这一处理方式对藏文的自动分词是很有必要的.

4)《词表》的调用　《词表》在运行统计软件的过程中,分别与汇总语料样本和7个分类语料样本先后被调用8次.

5)《词表》统计结果的排序　汇总语料和分类语料统计结果的降序排列,主要用于频率和累计频率以及通用度统计与分析;原序排列主要用于词条的人工检索与研究.

3　藏文字符、部件、音节、词汇频度统计

3.1　藏文字符统计

藏文字符(包括梵音字符)在北大方正藏文True Type字库中有4 483个;在华光藏文字库中有4 981个.现行藏文报刊、图书的电子文本大部分为方正,少部分为华光.华光文本需经转换软件处理为方正文本才能用于统计,所以统计出的字符仅限于方正藏文字库的字符系统以内,亦即均为方正系统可以屏显或打印出的字符.

藏文字符为全角字符,我们在程序设计时只把全角字符作为统计对象,而把半角字符(ASCII

码)经过滤预处理,既不作为统计对象,也不计入字符总数,这样就可以保证把文本中的所有藏文字符准确地检索出来．字符统计表中只列出藏文和梵音字符,藏文中所用的非藏文字符另外单列．

3.2 藏文部件统计

(1) 藏文部件 藏文部件是藏文字符最小的构成成分,共有4个元音,30个辅音(其中包括用作基字、前后加字、又后加字的辅音),3个上加字,5个下加字(包括ཨ་ཚ་ཡ།),还有藏文数字、标点符号和若干梵音部件及文化符．部件纵向叠加构成字符时,藏文字符最多有4层,梵音字符最多有7层．字符无论有几层,上平线始终保持不变．在电子文本中,任何一个藏文字符都是由部件构成的,每个部件各有其代码(内部码和交换码)．当部件构成字符后,就生成该字符的代码,除非某部件直接用作字符时才保留其自身的代码．计算机只能检索出直接用作字符的部件(部件即字符,字符即部件),而无法从两个以上部件组合的字符中去检索部件．因此,藏文部件的统计需要在字符统计的基础上,将字符按部件的分类进行归类后,从中再人工拆分统计出部件及其数据来．

(2) 藏文隔字符与隔句符统计

① 隔字符(ཚེག)是一个很特殊的字符．它仅占一般藏文字符字型宽度的1/8,但它仍然是一个全角字符．藏文的一个音节,相当于汉文一个字的读音．隔字符的使用频率是最高的,它约占藏文字符的33%．在统计分析藏文字符和部件时,需要视不同情况对隔字符加以区别对待:当对样本中的任意藏文字符作统计分析时,隔字符同样是其中的一个字符,应当客观地记录下它的频次并计算出其频率;当对构成音节的字符和构成字符的部件做统计分析时,因为它只起间隔音节的作用,如同汉文字与字之间不占字节的空隙一样,所以就不能把它计算在字符或部件的总数之中,这样才能真实地反映出字符或构件在文本中的实际情况,因而也就不能把隔字符的频率计入藏文字符和部件的累计频率．以往有关藏文字符和部件的统计研究,把隔字符、隔句符的频次计入了字符和部件的总数,并计入了累计频率,这样的统计结果自然会存在一定的误差．

② 隔句符(ཉྱིས།)在藏文中也是一个比较特殊的字符,它和隔字符所占的字形宽度一样,也是全角字符．隔句符在文本中有“单垂符”、“双垂符”和“四垂符”等用法,主要用作句读符号．

在藏文文本中,音节之间用隔字符间隔,句末用隔句符间隔,句首一般有空格,词汇之间没有自然的分词标记．当统计音节和词汇时,为了程序设计简便起见,需要把隔字符作为统一的检索识别标记,这就需要把句末与句首之间的隔句符和空格统统替换为隔字符,用隔字符把文本中的音节无缝链接起来．在对每一个分类语料库的隔句符作替换的过程中,我们从中抽取了一定数量的不同文体、风格的样本,记录下被替换隔句符的数据,然后依据这些样本的总字符数计算出隔句符在全部语料库中的数量及其频率.这样既解决了隔句符的数据统计问题,也解决了音节和词汇(相对《词表》而言)的识别标记问题,从而为音节和词汇的检索统计提供了整齐划一的识别条件．

(3) 数字及特殊字符统计

① 数字字符统计:藏文字符均为全角字符,而数字字符在藏文电子文本中既有全角状态下录入的,也有半角状态下录入的．为了便于统计,我们把藏文语料库电子文本中的所有ASCII码当作“无效字符”(Data invalid)在程序中做了过滤处理．这样统计出来的字符都是全角字符,其中也包括少量的全角数字字符,而在ASCII码状态下录入的半角数字字符则未被统计．然而,在藏文文本中无论是全角还是半角字符,数字字符是不可或缺的．为此,我们对“数字”又作了专门的统计和分析．

A. 将7个分类语料库的半角阿拉伯数字“替换”为全角阿拉伯数字之后,另行运行程序单独作了一次统计．

B. 将科技类和其他6个分类语料库的数字分别进行了统计与分析,相关数据参见下文和本课

题研究报告附录．

② 特殊字符统计　藏文文本中有一些较为常用，但又不属于藏文“书写符号系统”的特殊符号，如借用汉语或英语的引号、逗号、顿号、问号、中圆点号、省略号、感叹号、破折号、相至号、书名号、圆括号、方括号、尖括号、百分比号等符号．鉴于这些特殊符号在藏文文法中没有使用之规范，而在藏文真实文本中又在使用，所以在本课题研究正式的藏文字符、构件统计表中未列出，另行列于附录．

3.3　藏文音节统计

藏文的音节(ཚིག་སྡེབ་)是最小的不可划分的语素或单音节词，由 1 至 4 个字符横向构成．藏文的一个音节，就是由隔字符间隔的一个完整读音，相当于汉文一个字的读音．藏文的“音节频度统计”，相当于汉文的“字频统计”．藏文的“文字计量”与汉字相比，若以字符为单位则嫌小，以音节为单位又偏大，另外还要考虑隔字符和句读符号等因素．所以，这也是一个值得研究解决的问题．

对藏文进行音节的频度统计比较容易实现．当把统计样本处理为由隔字符无缝链接的音节串，只要以两个隔字符为界(‘～’)，以“自检”的方式进行检索、累计和记录，即把两个隔字符之间相同代码的字符或字符串及其频次按降序排列，便可准确统计出藏文音节的频次．

中文的字频统计对中文信息处理的许多领域都发挥了极为重要的作用，藏文的音节频度统计对藏文信息处理同样具有重要价值．本课题研究对汇总语料库音节、词汇和《词表》音节做了三种不同形式的统计．之所以要对《词表》所生成的文本作音节频度统计，是因为《词表》本身也是一个藏文“规范用词”的音节之集大成．动态语料库中的音节是开放式的，而《词表》中的音节相对而言是封闭的，用“封闭式”的音节对比分析“开放式”的音节，有助于对藏文音节的研究．从不同的层面和角度全面了解藏文语素的状况，尤其是其中使用频度高、构词能力强的语素对藏文识字教学和词汇学研究具有重要的参考价值．高频语素或词的集中识读，可起到“举一反三”甚至“以一当十”的学习效果．

3.4　藏文词汇频度统计

藏文和汉文一样，词与词在句中不象西文那样有明显的切分标记，因而藏文也存在难度很大的分词问题．由于现阶段藏语语料库的建设才刚刚起步，虽然已有一些很好的藏文分词策略、方法的研究成果，但尚无统一的藏文分词规范、标注集、通用自动分词软件和规模足够大的熟语料库，尤其是在藏文编码和字处理软件互不兼容的状况下，现有研究成果的技术和资源都无法实现共享，加之限于研究力量的极其薄弱，故基于数量级足够大的藏文语料库自动分词的词汇频度统计还有待时日．

鉴于本课题针对词汇的计量研究，主要是面向现代藏文(书面语)基本词汇和常用词汇的统计与分析．所以，我们采取了以“七分一总”语料库为统计样本，调用《词表》进行词汇匹配统计的方法．因此，《词表》的质量成为本课题研究质量保障的关键之所在．以下对《词表》和词汇频度统计的质量控制加以说明．

(1)《词表》的词条选自《藏汉大词典》等 8 部词典、正字法及其他专业词典共 34 141 条．《词表》的选词来源、范围、数量和覆盖率，客观上已达到最大限度，加之，词条的补遗及校对也是尽了最大努力．

(2)《词表》的选词标准，既充分考虑到藏文这种特殊文字的实际情况，也参照了一系列有关的规则、方法．本《词表》的选词具有一定的理论依据并借鉴了相关的实践经验．

(3)《信息处理用现代汉语常用词表》的一级常用词表为 6 994 条，二级常用词表为 29 970 条，单字词表为 3 522 条，其覆盖率在 98.5% 以上[10]．据本课题词汇统计结果表明：汇总语料库累计

频率达99%的藏文词汇量为9 000;《词表》中有14 526个词条的覆盖率仅占1%;频次为0次的词条有10 598条,占词条总数的31.057%.据马进武教授和其他几位藏族学者对其中的高频词、低频词和0词次的词汇进行分析,认为统计结果完全符合藏文使用的实际情况.1万多条罕见的0词次词汇,反映出藏文词典编纂以及文字规范中存在的一些问题.从统计结果中我们可以看出,《词表》的词汇总容量,是覆盖率达98.5%的9 000词的4倍,可见《词表》的词汇容量是足够的.由此可见,本《词表》的选词数量和覆盖率符合词频统计的需求.

(4) 我们用词汇频度统计结果生成的按原序排列的《词表》,对照《汉藏英对照常用词手册》[11]《中小学藏语文词语解释集》和《藏文正字法——语灯详释》[3]进行常用词汇对比分析时,复核出《词表》中有3个频率很低的词条有误,校对的差错率仅占词条总数的0.0009%.这说明《词表》校对的准确性是可靠的.

(5) 藏文音节的拼写是有规则的,是可以穷尽的.我们利用藏文音节的这一特性,用"封闭"的音节统计结果对照"开放"的词汇统计结果进行了对比分析,从而对藏文词汇的语素或词根及其构词规律有了比较清楚的了解.同时,据我们分析,藏文音节(即语素、词根、词缀、单音节词、语法虚词)的覆盖率为100%;书面语基本词汇和常用词汇(1~4个音节)的覆盖率可达99.9%以上;成语、熟语、合称词、缩略词、拟声词的覆盖率不低于98%.

综上所述,本《词表》的选词来源、范围、规则、校对和词汇的覆盖率以及频度统计的实现与正确性验证等一系列质量控制都是比较完善和可靠的.

4 藏文字符、音节、词汇频度统计的模型、算法与实现

4.1 提取语料库统计特征的数学模型

对藏文字符、音节、词汇频度(或叫频次)进行统计分析时,藏文语料库的规模及其结构是影响统计结果的重要因素.对于7个不同的语料分类,其词汇特征也不尽相同.因此,如何把不同类别藏文语料库的统计特征进行统一处理,从中提取出真实的频度统计信息,是藏文频度统计分析要解决的一个重要问题.在提取通用的统计特征时,必须对不同的分类语料库进行归一化处理.下面我们给出处理的理论分析:

定义1 设C_H是一人给定的语料库,$T_i(i=0,1,\cdots,n)$是定义的标记,$N(T_i)$是标记T_i在语料库C_H中出现的总次数,则定义语料库C_H的规模$D(C_H)$如下所示,它反映了语料库的大小:$D(C_H)=\sum_{i=0}^{n}N(T_i)$.

定义以下符号——M_g:通用语言模型;Ms_k:分类语料库专用语言模型,$k=0,1,\cdots,m$;T_i,T_j:两个给定的标记;$N(T_i)$:标记T_i在汇总语料库中出现的次数;$N_k(T_i)$:标记T_i在分类语料库S_k中出现的次数;P_i:标记T_i在汇总语料库中出现的概率.$P_i=\frac{N(T_i)}{D(M_g)}$,$P_i(S_k)$:标记$T_i$在分类语料库$S_k$中出现的概率.$P_i(S_k)=\frac{N_k(T_i)}{D(Ms_k)}$.

根据以上定义,把多个分类语料库合并成一个汇总语料库时:

$$P_i=\frac{\sum_{k=0}^{m}N_k(T_i)}{D(M_g)} \tag{1}$$

其中:$D(M_g)=\sum_{K=0}^{m}D(Ms_k)$.这实际上相当于将分类语料库混合在一起使用.这样做固然也可

以,但问题在于:第一,由于多个分类语料库的规模不一定完全一致,规模略大一点的语料库会占略大一点的比重;第二,由于有多个分类语料库,我们希望每个分类语料库只建立该分类语料库的统计模型,在使用时再与其他分类语料库模型动态合成,这样既便于对模型的管理和更新,又可以有效地节省存储空间;第三,既便是同一个语料分类也有不同的专业层次,在处理时应分别对待,但又不可能为每一个层次建立不同的模型.比较可行的办法是通过加权比重参数的调节以达到适应不同专业层次的问题,既对于专业性强的文本加大分类语料库模型的加权系数,反之则减小比重.

根据以上分析进行如下处理,其中的处理原则是:当忽略两个模型中的一个时,合成的结果与单独使用未被忽略时的模型一致.

首先,对统计结果进行规格化处理,使得式(1)的计算结果与分类语料库的规模无关,这样式(1)成为下式:

$$P_i = \frac{\sum_{k=0}^{m} P_i(S_k)}{m}. \qquad (2)$$

其次,为了考虑分类语料的层次性,对分类语料库的统计结果进行加权处理,式(2)成为下式:

$$P_i = \frac{\sum_{k=0}^{m} \alpha_k P_i(S_k)}{\sum_{k=0}^{m} \alpha_k}. \qquad (3)$$

其中 α_k 是针对第 k 个分类语料模型的调节系数,当 $\alpha_k > 1$ 时增强该分类语料模型的作用,当 $\alpha_k < 1$ 时减弱该分类语料模型的作用,α_k 取不同的值,可以在分类语料中不同的层次之间进行调节.

4.2 统计分析的数据结构

在藏文音节、词汇频度统计过程中,无论是对汇总语料库以"自检"的方式进行音节统计,还是调用机载藏文《词表》对汇总语料库和7个分类语料库进行词汇统计,主要的操作有两个:即藏文字符串的查找和字符串的更新(主要是字符串本身的累加和相应统计数据的实时修改).由于藏文语料库的规模和机载藏文《词表》的词条数目很大,如果数据结构设计不合理的话,就会导致查找深度太大,以至造成统计过程复杂度太大的问题.因此合理的数据结构设计是本课题研究能否实现的关键之所在.

在统计中我们主要使用了平衡二叉树(Balanced Binary Tree 或 height－balanced tree)结构(又称 AVL 树),以及与 trie 树的结合使用.下面对两种数据结构的设计分别加以说明.

(1) 平衡二叉树结构

① 平衡二叉树具有这样的性质:它是一个排序的二叉树结构,有左子树和右子树.

按照这样的性质构成的平衡二叉树,可以保证对于当前的语料库统计特征,其整个树结构的深度是所有可能树结构中最小的.

(2) 平衡二叉树的查找操作性能分析

假设藏文中含有 n 个不同的关键字,N_h 表示深度为 h 的平衡树中含有的最少结点数.显然,$N_0 = 0, N_1 = 1, N_2 = 2$,并且 $N_h = N_{h-1} + N_{h-2} + 1$.

利用数学归纳法证明:当 $h \geqslant 0$ 时,$N_h = F_{h+2} - 1$,而 F_h 约等于 $\phi^h / \sqrt{5}$(其中 $\phi^h = \frac{1+\sqrt{5}}{2}$),则 N_h 约等 $\phi^{h+2}/\sqrt{5} - 1$.反之,含有 n 个结点的平衡树的最大深度为 $\log_\phi(\sqrt{5}(n+1)) - 2$.因此,在平衡树上进行查找的时间复杂度为 $O(\log n)$.

3) Trie 树结构

除了上述采用平衡二叉树结构以保证数据查找的效率外，我们还采用了 trie 树这种优秀的字典方式数据压缩方法．设计思路可归纳为：假设有 4 个词条（为了便于说明起见，这里分别用符号表示为 abc，abce，abde，abdf），它们分别有一部分构件是相同的．按一般常规的做法，在统计时要分别记录这 4 个词条，这样至少要占用 15 个存储单元用于数据本身的存储．当利用 trie 树结构时，则可以让那些相同的构件部分只存储一次，这样仅需 7 个存储单元就可以记录这 4 个词条，几乎节省了相当一半的存储空间．

由于藏文是一种特殊的拼音文字，任何一个音节都是由 1～4 个字符拼合而成的，而藏文字符的数目并不大，因此，面对庞大的藏文语料数据库，采用 trie 树结构进行统计过程中的数据压缩，就能够使得数据所需要的存储空间大大减少，同时也避免了统计时所需内存空间不够的问题．

4.3　统计数据准确性的验驻

汇总语料库与 7 个分类语料库的藏文字符、音节、词汇频度统计完成后，经一系列文本转换操作，随机抽取了几百个频次为十几到几十、几百的字符、音节、词条，在其所属的语料库中用方正“查找”编辑功能进行了人工查找，以复核、验证统计结果的准确性，复核的结果表明准确率为 99％左右．另据推断，在句中因受隔字符的书写格式不规范或不统一的影响，某些词的统计准确率会受到略微影响．

5　藏文词汇频率、累计频率与通用度统计分析

［按：另见卢亚军、罗广专题撰文《藏文词汇通用度统计研究》一文，载《图书与情报》2006 年第 3 期 P74－77］

词汇的频度、频率、累计频率以及通用度，基本上可以比较全面、准确地反映一种语言文字中词汇使用状况的主要面貌，从而可以为词汇的应用研究提供基本的词目及其相关数据信息．以下我们对这一部分任务的实现加以论述．

5.1　总体任务

第一步，依据汇总语料库降序词条及其频次数据，计算词汇的频率和累计频率；第二步，使 7 个分类语料库的词条频次分别一一与汇总语料库相应词条对应，并计算出该词条的通用度；第三步，将运行“藏文词汇频率、累计频率与通用度统计软件”的统计结果，读入方正藏文字处理系统，进行表格编辑，完成《藏文词汇频率与通用度统计分析对照表》，最后进行正确性验证．

5.2　设计思路

根据总体任务的三个步骤，程序设计思路用文字表述为：

——先读取方正藏文系统下的汇总语料库词汇频次统计结果文件，从中分解出藏文词条（字符串）及其频次并存入列表；

——再分别读取 7 个分类语料库词汇频次统计结果文件，从中分解出藏文词条（字符串）及其频次并存入列表；

——依据汇总语料库词条（字符串）使 7 个分类语料库相应词条（字符串）的频次与之相对应，并按 BK（报刊类）、WX（文学类）、JY（教育类）、KJ（科技类）、FX（佛学类）、LS（历史类）、WM（五明类）的顺序，依次将词条频次存入列表.

——读取所有文件后，计算词条的频率、累计频率和通用度；

——以词汇的通用度降序排序；

——将统计结果（通用度降序排序）文件读入方正藏文系统，进行预设标记替换和超大规模表格编辑排版；对频率、累计频率和通用度统计结果进行抽查验算；对排版结果逐行进行逐页检查．

5.3 算法实现(在 Windows 环境下用 Delphi 实现)

(1) 读取文本文件 藏文词汇频次统计结果文本文件不像数据文件那样有字段结构,需要逐个分析出词条(字符串)和频次数据.在 16 进制码下只能看到藏文词条(字符串)和频次数据之间有特殊字符作为分隔符和不等量的空格,只要取得分隔符就可以将藏文词条和频次数据分拣出来.为此,需要设立一个文件指针 pFile 总是指向文件中的当前字符.在判断 pFile 所指的字符时,如果是藏文词汇的词头(十六进制显示为 $ c0),则将当前模式设为词汇模式(变量 isword:=true),在词汇模式下 pFile 所指的内容将存入词组 cizu[ ],然后 pFile 加 1 指向下一个字符;如果 pFile 所指的内容不是分隔符而是数字字符,则将当前模式设为数字模式(变量 isword:=false),在数字模式下,pFile 所指内容存入数组 shuzu[ ].这时 pFile 继续向下移动,如果 pFile 指向换行符,则将词组 cizu[ ]、shuzu[ ]存入列表并清零;接着继续移动 pFil 指向下一个藏文词条.如此循环直止读完所有文件.

2) 分拣词汇频次统计数据 当汇总语料库词汇频次统计结果分拣完后,再依次读入 7 个分类语料库词汇的频次统计结果文件,并且要求分类语料库统计结果的词汇频次与汇总语料库统计结果的相应词汇要一一对应并按顺序跟进排列.假设词条 C5 在文件 FZ(汇总语料库词条)的位置是第 5 行,而同一词条在文件 BK(报刊分类语料库词条)的位置是第 16 行,那么,当读取 FZ 后,FZ 作为基准的第 1 个文件列表为新建列表,词条 C5 存入第 5 行,而在读取文件 BK 中的词条 C5 后,必须查找 FZ 的 C5 在列表中的位置,并将其频次存入与之列表相同的位置.

按照一般的查询方法,即每遇到一个词条 Cx,就从列表开始处查找,直到找到该词条.这种方法速度太慢,尤其当词条 Cx 处于列表尾部时,程序将执行许多空循环.经仔细分析,我们发现词汇的频次高低,无论哪个分类语料库的词汇都与汇总语料库相应词汇的频次高低有一个对应的范围.这就使我们产生了用"设置查找指针"办法解决这一难题的思路.这样,随着查找指针逐渐向下移动,被查找词条和循环次数递减,查找速度随之显著提高.

(3) 统计频率和累计频率与通用度 本课题研究考察、分析了国内有关词频统计的多种方法,如《现代汉语词频词典》[12]采用的"使用度"概念和公式,《信息处理用现代汉语常用词词表》提出并采用的"双选词函数"概念和公式,还有其他一些概念和方法."这些方法所用到的公式不但非常复杂,而且所得到的结果并不十分理想."[13]词汇的"通用度"是词频统计的一种新概念和新方法.原北京语言学院和中国社科院语言文字研究所"报刊新闻词语的统计与分析"课题组,从实践的角度进一步证实了通用度计算公式的合理性和实用性.本课题研究认为"通用度"的方法比较理想,故予采用之.

① 词汇的频率和累计频率统计

词汇的频度(频次或词次)指某一词在统计样本中出现或使用的次数;频率指某一词条的频度占全部词汇频度累计总数的比例.频率计算公式为:词汇频率=某词频度/总词汇频度(%);累计频率指某一词条的累计频度占统计样本总频度的比例.累计频率计算公式为:累计频率=上一个词条的累计频率+本词条的频率(%).

汇总语料库词汇的频率和累计频率,由程序统计出汇总语料库的总词汇频度后自动完成统计.随后人工进行抽样检验.

② 词汇的通用度计算

通用度计算公式[13]为: $T=(\sqrt{n1}+\sqrt{n2}+\cdots+\sqrt{nk})^2/k$.

分类统计样本为 k 组,每组样本的数量大致相等.某一个词条在 $1,2,\cdots,k$ 组的出现次数分别为 $n1,n2,\cdots,nk$,那么这个词条的通用度为 T(通用度算法的具体实现此略).

(4) 保存统计结果文件　将列表中的统计与分析结果以文本文件格式输出,以便在方正藏文字处理系统中编辑、排版.

5.4　制作《藏文词汇频度与通用度统计分析对照表》

将藏文词汇频度与通用度统计分析结果文件,读入方正藏文字处理系统进行一系列标记替换、编辑和排版操作之后,还要对词汇频率、累计频率与通用度的正确性进行抽样验算.该统计分析对照表共有词条数为23 580条,总词次为11 159 150,字数约180万字.

6　藏文字符、部件、音节、词汇频度与通用度统计数据分析

6.1　有关基本参数

(1) 汇总语料库(统计样本)字节数为70 040 328(bytes),全角字符数为34 070 212(其中包括非藏文字符,不包括ASCII码),ASCII码为1 899 905(bytes).

(2) 隔字符数为1个,频次为10 471 235;隔句符数为1个,频次为872 878.

(3) 藏文字符为1 578个(其中包括隔句符、蛇形垂符、聚宝垂符和梵音字符,不包括隔字符和藏文中使用的汉英标点及计算符号、阿拉伯数字等非藏文字符,以下称"纯藏文字符"),频次为23 067 300;藏文字符与隔字符的合计频次为33 538 535.

(4) 藏文部件数为78个(包括常用的梵音部件,不包括隔字符和阿拉伯数字、藏文中使用的汉英标点及计算符号),频次为33 466 640.

(5) 汇总语料库中检出的藏文音节数为7 976个(不包括拼写错误及非藏文字符串),频次为11 423 207.从《词表》中检出的藏文音节数为5201个,频次为73836.

(6)《词表》的词汇容量为34 141条.

(7) 汇总语料库中检出词条23 580条,词汇总频次为11 159 150;

7个分类语料库(统计样本)字节显示及检出词汇:

分　类	data read (全角)	data invalid (半角)	词　条
报刊类	5 003 053	228 200	14 558
文学类	5 002 318	235 790	14 768
教育类	5 003 279	233 704	15 059
科技类	5 004 830	589 721	9 886
佛学类	5 004 743	204 087	12 545
历史类	5 000 878	201 143	15 295
五明类	5 001 150	207 262	15 827

(8) 藏文中使用的非藏文字符个数为34个(包括全、半角阿拉伯数字、汉英标点及计算符号和ASCII码),频次为531 677,约占总字符数的1.56%.

6.2　有关数据分析

(1) 藏文隔字符约占汇总语料库全角字符总数的30.7%;隔句符约占2.56%.

(2) 纯藏文字符的频次约占全角字符总数的67.7%;纯藏文字符和隔字符合计共约占98.44%.

(3) 实际检出的纯藏文字符个数为1 578个.这1 578个藏文字符的前301个字符,累计频率为99.2%,前500个字符达到99.915%,而其余1079个字符仅占0.085%.至于方正、华光藏文字库中的其他2 900多个字符能占到0.085%中的多少,乃可想而知.检出0词次的藏文字符34个,均为字库中"张冠李戴"的错误拼写字符.

从本课题研究藏文字符统计与数据分析的结果中,我们可以确切地得知——现代藏文的内码、

字模、字库、输入法设计、文字识别以及互联网和通讯等藏文信息处理领域所需的藏文字符的样式和数目为601个左右即足够使用．至于其他一切“开放式”梵音字符的问题，可通过叠加组合的方式解决．

(4) 按藏文部件统计结果的降序排列，前10个部件的累计频率为57.48%，第20个为85.8%，第30为95.75%，第40已达99.59%，截止到第56则高达99.95%．梵音部件的频率均在0.0011%以下．藏文部件实现动态叠接组合，有五六十个部件就可达到99.95%以上的覆盖率．藏文部件、字符的统计数据，对藏文编码标准的研究和计算机键盘布局及输入法设计乃是最重要的数据依据．

(5) 藏文字符平均由1.5129个部件构成．藏文部件频次除以纯藏文字符频次(其中不包括藏文隔字符、隔句符、蛇形符和聚宝符)．

(6) 藏文音节平均由2.78个部件构成．按统计《大藏经》部件构成音节的算法[14]，“由隔字符与隔句符的频率之和的例数减1”，其计算结果为2．54个部件．根据我们的统计数据计算的结果则是2.78个部件，即1 000÷(238.3191+26.082)－1=2.78．据分析，二者计算结果的不同是因为各自所采用的统计样本不同，前者是以韵文为主，后者是以散文为主．此外，前者把隔字符也计入了部件中．

(7) 藏文音节平均由两个字符构成．纯藏文字符(其中不包括隔字符和隔句符、蛇形垂符、聚宝垂符等句读符号)频次除以藏文音节频次，得数为1.9428．此外，要计算纯藏文电子文本中的音节，隔字符等句读符号就要计算在内，算法是字节数/2÷2.9825(2.9825这个数值是我们测算普通藏文文本字符频次与音节频次的比值)．有研究者以《藏汉对照拉萨口语词典》《拉萨口语读本》词汇表和《藏语简志》词汇表为统计样本，得出“藏字(指藏文音节)的平均字符长度为3.6780字符．”[15]这一统计分析结果显然只能对其统计样本——词典、词汇表而言是准确的．因为词典中有不少“三时一式”的动词形态是由3个和4个字符构成的，而在真实文本中大多数音节是由两个和一个字符构成的．

(8) 由于藏文音节在真实文本中是两个隔字符及其与隔句符、空格之间的字符或字符串，所以从理论上说，隔字符数+1即为音节数．实际上我们的音节频次统计结果与隔字符(包括藏文句读符号)频次之比的误差仅为0.068%，而这些误差是由真实文本中存在的书写不规范、拼写错误、长腿字符处在句末，以及非藏文字符串和极少数不能识别的梵音字符串所致．统计结果中有这样一个微量值的误差也是客观存在的反映，所以说我们以“自检”的方式所进行的音节统计数据是完全准确的．

(9) 汇总语料库中检出的前1 333音节的累计频率为95%，到2 787时已达99%；从“封闭式”的《词表》中检出的音节为5 201个．其中，前530个音节的累计频率为60%，1 200个音节的累计频率为80%，2 717个音节的累计频率为95%，3 666个音节的累计频率为98%，频度为1的音节有1 535个．这说明藏文的单音节，即语素或词根的构词能力很强．汉文是“3 000～4 000个常用字覆盖率达99.9%”[16]．依据藏文音节统计数据与西文和汉文作比较，可以确信藏文也属于“二次构词”文字，并有其独特的优势——兼备拼音文字和象形文字二者的优点．

(10) 汇总语料库中检出频次为1次以上的藏文词汇有23 580条，其中：第1 000条的频次为1 543，累计频率83.37%；第2 000条的频次为500，累计频率91.15%；第3 000条的频次为251，累计频率94.31%；第5 000条的频次为148，累计频率96.02%；第9 000条的频次为28，累计频率99.%；其余14 526个词条的覆盖率仅占1%．频次为0次的词条有10 598条，占《词表》词条总数的31.057%．据有关专家研究，“当今所有主要语种日常实际使用的词大致保持在5万个上下．实

行一次构词为主、'字话一律'的英、法、德、俄等语种,其书面词形就是5万多个字;而以二次构词为主的汉语,其书面语只由4 000～5 000个汉字组成,是英文的1/10,是俄文的1/12."[17]. 略懂一点藏文的人也能感觉到,同样的内容,藏文图书比汉文图书差不多要厚一倍. 统计结果也表明,汉文书面语大约是藏文书面语(汉译藏)的1/2. 其实,正是因为藏文具有比较丰富的语法虚词和"三时一式"的动词形态,所以才比以语序为主而缺乏语法虚词和动词形态的汉文多了一些具有语法功能的句子成分. 这点"多"非但不是藏文的缺点,反而是其独俱特色的优点.

(11) 数字在任何一种语言中都是必不可少的. 但数字在藏文电子文本中呈现出比较复杂的情形. 一方面由于文体的不同,数字的使用频率也不尽相同. 如果把10个阿拉伯数字计入藏文字符,就会干扰整个藏文字符体系自身的规律. 另一方面,数字在藏文电子文本中,有藏文大小写,有阿拉伯数字,有少量在不同输入法下录入的全角数字,也有许多ASCII码下录入的半角数字. 如此五花八门的状况都是各种人为因素造成的,实际上数字在藏文中的使用还是有其规律的.

为了准确统计并客观反映数字在藏文中的使用规律,我们对阿拉伯数字(不包括藏文数字)进行全半角字符的转换处理之后另行作了统计. 统计结果表明:10个阿拉伯数字在科技分类语料库中的平均频次为15 658,频率为4.46‰;在其他6个分类语料库中的平均频次为13 075,频率为3.73‰.

(12) 藏文词汇通用度的统计与分析,是本课题研究对词汇频度统计新概念、新方法的一次大规模的应用. 正因为如此,我们对藏文词汇使用状况的了解和信息处理用词汇分级以及教学、扫盲用词汇的选择有了科学的依据和确切的把握.

以上统计数据和相关分析,我们都以人工的方式进行了实测比对、验证,未发现有出乎意料的误差. 而真正出乎意料的是——我们发现藏文是世界上惟一由中国55个少数民族之一的藏族所独创的一种兼备拼音文字和象形文字二者优点的文字.

总之,本课题研究语料库规模的足够大,统计样本的合理化,《词表》的覆盖率高,统计软件的适用性好,统计结果的准确性和可信度高,这一系列重要因素共同构成了本课题研究层层关联,环环相扣的质量体系.

7 本课题应用研究及取得的阶段性成果

7.1 本课题研究开发的应用软件:

① 华光－方正、方正－华光藏文转换软件;② 藏文字符频度统计软件;③ 藏文音节频度统计软件;④ 藏文词汇频度统计软件;⑤ 藏文音节频率、累计频率统计软件;⑥ 藏文词汇频率、累计频率与通用度统计分析软件.

7.2 本课题应用研究及取得的阶段性成果:

① 藏文部件频度统计表;② 藏文部件频度统计降序表;③ 藏文元辅音与基字及上下加字字符频度分类统计表;④ 藏文字符频度统计表;⑤ 藏文音节频率(语料库)统计表;⑥ 藏文音节频率(词典)统计表;⑦《藏文词汇频率词目》;⑧《藏文词汇频率与通用度统计分析对照表》;⑨ 大型藏文语料库数据库(1.3亿字节);⑩ 机载电子藏文《词表》(34 141个词条).

参考文献:

[1] 黄伯荣,廖序东. 现代汉语. (2版增订本)[M]. 北京:高等教育出版社,1997,162.

[2] 张怡菘. 藏汉大辞典[Z]. 北京:民族出版社,1985,7.

[3] 7种藏文词典和正字法分别为: