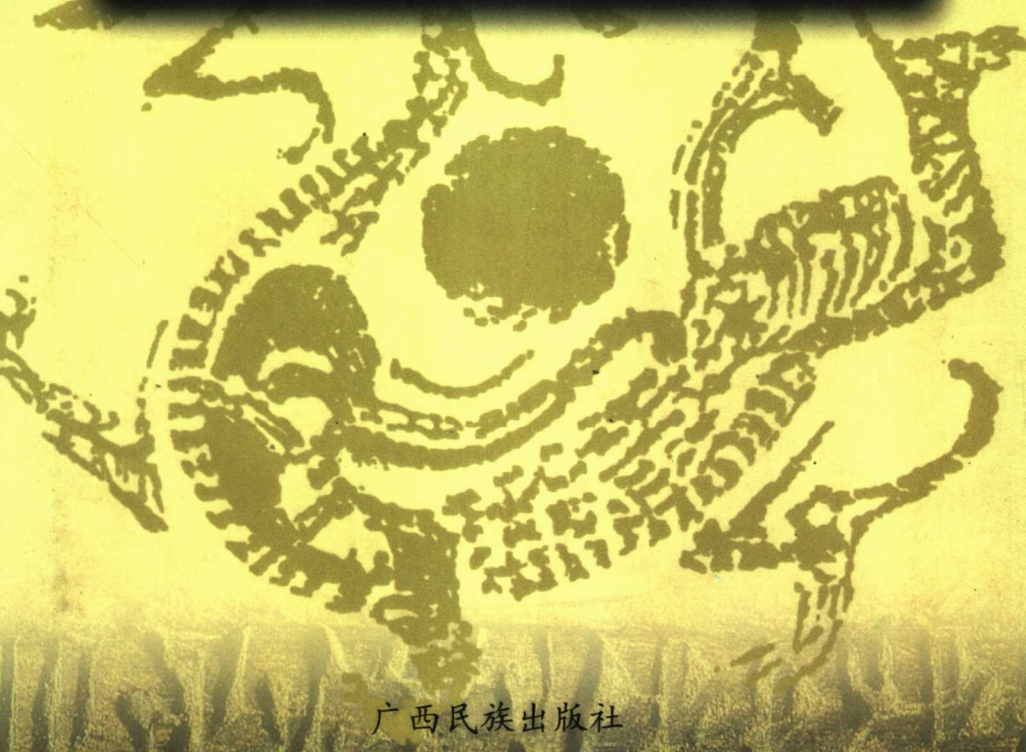
Cognate Words in Sino-Tibetan Languages

汉藏语同源词研究（三）

汉藏语研究的方法论探索

丁邦新　孙宏开　主编



广西民族出版社

汉藏语同源词研究（三）

——汉藏语研究的方法论探索

丁邦新　孙宏开　主编

广西民族出版社

图书在版编目（CIP）数据

汉藏语同源词研究（三）：汉藏语研究的方法论探索/丁邦新，孙宏开主编. —南宁：广西民族出版社，2004.6
ISBN 7-5363-4761-8

Ⅰ.汉… Ⅱ.①丁…②孙… Ⅲ.同源词－比较词汇学－藏语、汉语 Ⅳ.H214.3

中国版本图书馆 CIP 数据核字（2004）第 056594 号

汉藏语同源词研究（三）

——汉藏语研究的方法论探索

丁邦新　孙宏开　主编

责任编辑 韦光化
封面设计 陈　卓
责任校对 韦彩娟　卢芳芳　黄一清
责任印刷 姜为民
出版发行 广西民族出版社
印　　刷 广西区计委印刷厂
开　　本 850×1168　1/32
印　　张 17.25
字　　数 400 千
版　　次 2004 年 7 月第 1 版
印　　次 2004 年 7 月第 1 次印刷
印　　数 1—1000 册

ISBN 7-5363-4761-8/C・215　　定价：85.00 元

编者的话

在第一卷里，我们梳理了自 18 世纪以来汉藏语系的研究情况。汉藏语系虽然提出很早，也开展了一定范围的研究，但在研究方法方面基本上沿用的是传统的历史比较语言学的方法，这种方法使用在印欧语系方面取得了一定的成功，但汉藏语系语言文献少，情况复杂，语言之间的接触、影响十分久远而且深刻，地域性趋同十分明显。因此语言与语言之间的类型一致是同源关系还是接触关系难以判断，产生了很多分歧。在这种情况下，除了使用历史比较法外，人们在寻找新的适合于汉藏语系同源关系的研究方法。收入本书的几篇报告，就是作者在这方面进行的有益探索。

计算机的广泛运用是 20 世纪科学革命的一项重要体现，语言学科也基本上摆脱了过去手工排卡片的历史，并采用了一些新的思路、新的角度或新的方法来处理语言资料，观察语言现象。黄行的文章“汉藏语言关系的计量分析”是在课题组初步完成了《汉藏语同源词研究·词汇语音数据库》以后，在数据库基本资料和标记的基础上用计算机进行语言历史比较研究的一个尝试，对语言的同源关系和语音对应关系在过去仅仅做定性分析的基础上，开展定量的统计分析，以求得更科学的结论，进一步检验定性分析结论的可靠性。他从以下 3 个方面进行了探索：

1. 语音对应规律的概率统计。

2. 关系词的词阶分布。

3. 语言亲疏关系的计量分类。

陈保亚 1994 年完成了他的博士论文《论语言接触与语言联盟——汉越（侗台）语源关系的解释》（1996 年 7 月语文出版社出版）。他以斯瓦迪士（M. Swadesh）200 核心词作主要研究对象，把它分为第 100 词集和第 200 词集两个不同的阶。通过对大量材料的分析统计，他发现在语言接触和语言分化中，有语音对应的关系词的分布是对立的。在语言接触中，第 100 词集的关系词低于第 200 词集的关系词，在语言分化中情况正好相反。他进一步考察了汉语和侗台语之间一批对应比较严格的古代关系词的分布，发现第 100 词集的关系词远远低于第 200 词集的关系词，由此认为这一部分关系词不是侗台语和汉语有同源关系的证据，而是接触的证据，由于这批关系词数量较大，用语言联盟解释这种密切接触关系比较合理。著作出版后，国内外学者提出了不同的意见，主要是：

1. 汉台关系词在两阶核心词中的分布情况是否可以概括其他关系词的分布情况？如果统计不限制在 200 核心词，而是同一时间层面的全部关系词，有阶分析怎样进行？

2. 分析和寻找语音对应规则的关系是什么？为什么有的学者研究汉台关系词的结果是第 100 词中的关系词比例高于第 200 词的关系词比例，而另一些学者正好相反？

3. 早期关系词可以分成不同的层面，哪些层面的关系词是最重要的层面。

陈保亚、何方的《汉台核心一致对应语素的有阶分析》通过进一步研究回答了上述问题。重点放在对应语素时间层次的区分、核心一致对应的重要性、相对有阶分析、语素类聚有阶分析等问题上。

邢公畹先生在20世纪90年代提出了确认汉藏语系语言同源词的一种方法，叫“深层语义对应”比较法，1999年出版了《汉台语比较手册》，使用语义学比较法论证了汉语和侗台语的发生学关系。他先后共举出汉台语之间的305组深层语义对应例证，包含了585对汉台语同源词。在此期间他还发表《汉苗语语义学比较法试探研究》。

自1997年开始，邢凯就协助邢公畹从事语义学比较法的基础论证工作，先后发表了《历史比较法是建立语言史的有效工具》(1997)、《语义学比较法的逻辑基础》(2001)、《有关语义学比较法的理论问题》(2002)。文章重申了历史比较法的基本原则，认为语义学比较法是汉藏语系历史比较研究的新发展，它和“形态学比较法”有异曲同工之妙，是在传统历史比较法的形态学原则和语音学原则之外又提出了第三个原则——语义学原则。

收入本书的《语义学比较法》是作者对该项研究的总结。共分为三章：第一章语义学比较法的提出及理论基础，联系形态学比较法和语音学比较法，比较全面、深刻地阐明了语义学比较法的理论基础，并提出了该方法的六种辅助证明手段。在该文附录中把邢公畹举出的305组深层对应例证列成一个“汉语台语深层语义对应总表”。第二章归纳了16种深层对应的类型，补充了新例证，并使用该方法论证了汉台的同源关系。第三章是关于语义学比较法理论概括，并回答了学术界提出的质疑。

本课题的一项重大成果是建成了一个有利于汉藏语系历史比较研究的130种语言或方言和11部词典的词汇语音数据库。其中包括古汉语和汉语方言、侗台语、苗瑶语、藏缅语、南岛语、南亚语的词汇材料，各语族构拟和上古汉语构拟资料，每个语言或方言包括1332个基本词。每个词带40多种标记，包括音节数；每个音节的声、韵、调切分，有利于语音对应比较和同源词的确认；有词根位置的标注和词缀的标注；有词源的说明；词典

部分有利于做词族研究和语音对应的研究。

江荻博士用了近 5 年的时间完成了数据库设计和建设。本书收入了他对数据库基本内容的介绍和说明，以便读者了解和使用，把汉藏语系研究引向深入。他的文章包括：（1）“汉藏语同源词研究”数据库检索系统需求分析；（2）“汉藏语同源词研究”·数据库结构系统设计；（3）“汉藏语同源词研究”·语言代码表设计；（4）“汉藏语同源词研究”·大数据词典设计；（5）“汉藏语同源词研究”·数据库检索软件设计；（6）“汉藏语同源词研究”·属性代码设计；（7）“汉藏语同源词研究”·检索系统的具体实现；（8）“汉藏语同源词研究”·IPAJ 国际音标系统的设计。本系统具有强大的管理和检索功能。

1. 检索：其中有语言检索，指定某个或某些语言进行检索，语义类别检索：指定某语义类别检索；词条检索，指定词条（包括中文、英文、任一种民族语言词条）；数据结构元素检索〔包括输入指定的：1）音节，包括指定核心音节（核心音节具有属性标记）；2）声母；3）韵母；4）声调；5）前置辅音；6）基本辅音；7）后置辅音；8）元音；9）韵尾；10）语法描述属性；11）语音描述属性；11）数据组合检索；12）用户可选择词条筛选或浏览操作。

2. 输出：检索的结果允许用户制作成文本文件（TXT/W(ORD) 输出/存储；检索的结果可打印输出或制成表格方式输出；允许用户指定语言或数据输出的排列顺序。

3. 统计：对任一种检索结果均产生统计数据。

4. 输入和帮助：系统提供课题项目说明帮助；提供每种语言的简介信息帮助；提供每种语言的音系介绍，帮助用户了解语言情况；除了中英文词条及音标类检索需键盘输入外，其他中英文均尽量提供免录入方式输入；提供实时弹出的国际音标活动键盘图。

5. 系统维护：允许管理员进行部分数据修改、增加、删除；要求用户注册（用户名、密码、姓名、单位、通信地址），允许撤消注册；要求用户登录核对用户名和密码；管理员进行用户档案管理（增加和删除用户）；管理员密码维护（修改、更新）；允许用户进行密码设置和修改；用户使用记录：检索/查询、清除、备份；系统设置备份功能和紊乱恢复和纠错功能（定期数据备份和系统重索引）；用户管理：完成 STCP 系统用户的增加、修改和删除操作，规定用户的权限；日志管理：对用户使用情况进行登记，以备审计的需要。

我们一贯强调，汉藏语系历史比较研究是一种探索，无论是历史比较研究本身还是方法论的探索，都是这样。在这种情况下，完全取得一致意见是非常困难的，因此一定要坚持百花齐放和百家争鸣的方针，允许发表不同意见，我们是本着上述原则进行本丛书的编辑工作的。我们再次重申，收入各文集的专题讨论，仅仅代表作者的观点，并不代表编者的意见，编者和作者之间对于一些问题的看法并不完全一致。

最后，我们希望借此机会，再次感谢广西民族出版社的领导和编辑，他们在财力非常紧张的情况下，高质量地出版了这套丛书，在国内外学术界产生了良好的影响。

丁邦新　孙宏开

2004 年 1 月 31 日

目　录

汉藏语言关系的计量分析

黄　行

中国社会科学院民族学与人类学研究所

汉藏语系的语言数量多、分布区域广、结构复杂，特别是因为缺少历史文献记录，汉藏语言的发生学关系从语言资料和研究方法上一直没有得到充分的证实。《汉藏语同源词研究》建立了包括125个点的汉语、藏缅语、侗台语、苗瑶语及南亚、南岛诸语言方言（包括古代语言构拟）的基本词汇库，为研究汉藏语系的语言关系提供了可用于基础比较研究的详备语料。

汉藏语言关系的研究主要是以同源词研究为核心，其目的在于通过有语音对应关系的同源词系统建立诸汉藏语言的谱系关系。在语系层次上，汉藏诸语的关系相当疏远，语言之间同源词数量很少，语音对应关系也不很整齐。本尼迪克特—马提索夫的《汉藏语概论》（剑桥大学出版社1972）、柯蔚南先生的《汉藏语词汇比较手册》（1986）、俞敏先生的《汉藏同源字谱稿》（《民族语文》1989年1、2期）、全广镇先生的《汉藏语同源词综探》（台湾学生书局1996）、邢公畹先生的《汉藏语同源词初探》和陈其光先生的《汉语苗瑶语比较研究》（2001）等人的研究分别找到一些汉语和某些汉藏语系民族语言的同源词（或关系词），

并做了语音对应关系的比较和构拟。语族内部的语言关系要密切得多，因此同源词的数量颇大，语音对应也比较清楚。李方桂先生的《台语比较手册》（1977）和梁敏、张均如先生的《侗台语概论》（1996），王辅世先生的《苗语古音构拟》（东京外国语大学 1994），王辅世、毛宗武先生的《苗瑶语古音构拟》（1995）分别建立了比较系统的侗台语和苗瑶语同源词的语音对应关系和原始语言的古音构拟。

有语音对应关系的词不仅包括同源词，也包括语言长期接触吸收的借词，二者被合称为关系词。就关系词的研究而言，近年来学界比较注意关系词的核心层次（词阶）和年代层次的差别，并以此为切入点，试图通过关系词的核心程度的阶，以及关系词产生的年代来进一步区分同源词和借词。与汉藏语语言关系研究相关的另一个问题是语言关系的亲疏程度。所谓语言关系的亲疏一般是指语言结构单位的音值和类别的接近程度，这种接近程度可能主要取决于语言的发生学关系，但是也会受其他因素的影响。

汉藏语同源词或关系词的研究都与词汇的数量分布特性有直接的关系，因此本文拟用汉藏语同源词语料库和一些已有的语料库为样本，用侧重数量分析的方法对语言之间的语音对应规律、关系词和核心词的语源关系，以及语言的亲疏关系等有关的问题做一些分析方法上的探讨。

一、语音对应规律的概率统计

汉藏语比较研究各家的体系确定有语音对应关系的同源词所依据的标准并不完全一致。在语族层次，特别是内部关系相对比

较清楚的侗台语族和苗瑶语族，由于所掌握的语言方言点分布面广、代表性强，用于比较的词汇量也比较大，当然更主要的是这些语言具有显而易见的同源成分，因此同源词的语音对应规律是严格的；而对于那些关系比较疏远的语族以上层次的语言来说，同源词的语音对应规律就不是那么严格，一般是用“貌似”，即语音和词义比较相近的标准来确定同源词。这样处理的原因首先是因为语言关系疏远而不可能列举出数量较多、对应严格的同源词，但是如果按照经典的历史比较法，没有严格语音对应关系的词是否同源词是有疑问的。

所谓语音对应规律和其他自然社会规律一样是指在同样条件下可以重复出现的现象，因此语音对应规律是可以用概率统计证明的。陈保亚先生的《语言接触与语言联盟》（1996）专门用一节讨论语音对应规律的概率基础问题，指出语音对应规律是基于音类和词的语音对应的数量关系。由于任何语言的语音单位是有限的，而词汇的数量却很大，因此每个音类会随机地分布于一定数量的词当中，这种词和语音的随机对应不但不能作为语音对应的根据，反而是应当被排除的。举例来说，两种各有 50 个音类和 5 万个词的语言，它们之间任意两个音类的随机对应概率平均为 $1/50^2=1/2500$。换句话说，即使没有任何关系，这两种语言之间任意两个音类平均也会有 20 对词、全部词汇会有 1000 对词出现对应关系，这种对应显然不是同源关系造成的。只有当词的语音对应概率大大高于其随机对应概率的时候，才可以认为这种对应是有规律的或有原因的。陈著首次用概率统计的方法来描述汉藏语语音对应关系，并用具体的语料对这种方法做了验证，是值得称道的。但是陈的方法忽略了一个非常重要的语言现象，即任何语言的语音都会因音类标记性程度的差别而呈非均衡分布状态，即无标记音类出现频率高，有标记音类出现频率低。因此用平均随机概率去检验音类的分布概率是不合适的。本文在尝试对