几种南亚语的词源统计分析

高永奇

[提要]本文运用词源统计分析方法对我国的几种南亚语系语言进行分析，并通过无根树和在此基础上生成的有根树的比较分析，讨论这几种语言的系属分类情况及有关的其他问题。

一 我国南亚语系语言的分类及使用情况

南亚语系（Austro-Asiatic Family）是1907年由德国人类学家W·施密特（W·Schmidt）提出来的。他首先将这一语系的语言分为三大语族，其后又分成四个语族。后来，美国学者迪福乐将南亚语系分为三大语族：（1）蒙达语族；（2）尼科巴语族；（3）孟高棉语族。我国学者李道勇（1985）把南亚语系分为三大语族：（1）孟高棉语族；（2）扪达语族；（3）石芒沙孟语族。颜其香、周植志（1995）将南亚语系分为四大语族：（1）孟高棉语族；（2）越芒语族；（3）蒙达语族；（4）尼科巴语族。其他学者或者用三分的观点，或者四分，名称也不尽相同。这些学者的分类依据为：基本词汇数目、语音对应规律、语音面貌、语法特征的相似性，并根据基本词汇的相同相近的数量用来确定语言关系的远近。

持语系三分观点的学者都把我国的南亚语系语言归入孟高棉语族，持四分观点的学者则把我国的南亚语系语言分为孟高棉语族和越芒语族两类。下面是我国一些南亚语系语言采用语族四分的分布情况（颜其香、周植志1995的分类）：

- 越芒语族
  - 越芒语支——京语；
  - 巴琉语支——巴琉语（侬语）；
  - 莽语支——莽语；
- 孟高棉语族
  - 佤德昂语支——佤语、布朗语、德昂语等；
  - 格木语支——格木语（克木语）等；
  - ……
  - （九个语支）

二 词源统计分析方法

词源统计分析方法是借用生物学上关于物种进化关系分析的方法来分析语言的亲缘关系的一种计算统计方法。其理论基础就是有亲属关系的语言在演化过程中，其基本词汇的演变转化程度不同。邓晓华、王士元（2003a、2003b）曾采用这种方法分析苗瑶、藏缅语族语言的

亲缘关系和分化情况。词源统计分析不仅可以显示各种语言的亲疏关系，更可以显示出语言之间的亲缘距离。关于这种方法的具体情况，邓、王在其论文中有详细的说明，这里我们不再赘述。有两点需要说明：

1．选词范围与同源判断

词源统计分析的基础和前提是核心同源词的选取，选词上以 Swadesh 的 100 词表为标准。我们在分析我国南亚语系的几种语言时，同样以此为基础，个别从后 100 核心词中选取相应的词语补上。

词语比较时同源词身份的确定并不是一件容易的事。这是因为，第一，多数情况下我们不能明确区别是同源词还是借词，尽管我们可以制定相应的判别标准，但实际上仍会有一定数量的借词掺杂于其中。邓、王采用的是“分词目计算，即采用较严格的语义对应原则。”（2003a：256）同时，他们也指出：“如何正确地区别开借词和同源词，这取决于我们的历史音韵学知识体系以及对整个东亚区域的相关民族的历史文化及其变迁的理解。”（2003a：258）我们为了使统计的结果尽量接近事实，采用多语言并比的方法，即在几种语言共同参与下的比较。因为有些词语在两种语言的对比中很难分辨其间的关系，而在其他语言的参照下，其间的语音对应关系则可以显现。第二，由于比较的语言的数量总是有限的，几种参加比较的语言中很可能存在一些共同的外来词。在没有跟其他语言比较的时候，这些共同借词就很难被发现。这类共同借词并不影响在参加比较语言之间的关系，但对它们的语言分化年代的计算却有直接的影响。也就是说，共同借词对比较语言间的亲疏关系没有直接影响，但对亲缘关系的分析却有直接的影响。

2．参加比较的语言的数量

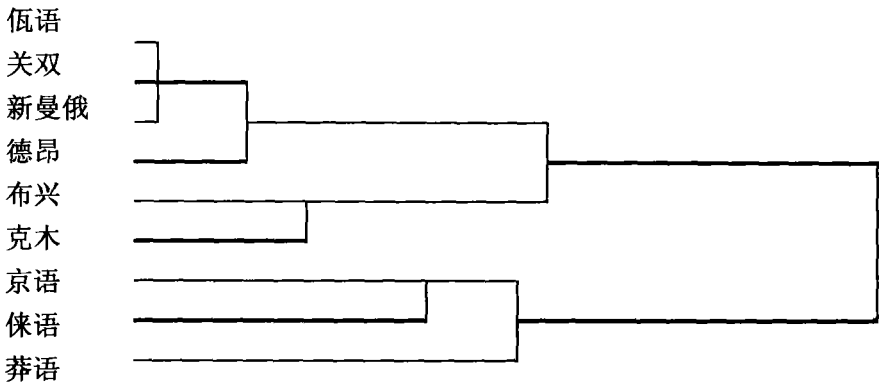
参加比较的语言的数量的多少直接影响到比较的结果。词源统计分析是对语言之间的分化距离进行的数理分析，参加比较的语言有的关系较近，有的关系则较远。最理想的做法是把一个语族或语支的所有语言都拿来进行对比，分析它们之间的相互关系。但事实上，在研究的开始阶段，我们并不能全面掌握整个语族、语支的情况，对其中每种语言的地位、相互关系往往要在综合分析之后才有结论。而词源统计分析并不依赖现有语言的结论。在部分语言参加统计分析的时候，词源统计分析的结果所反映的只是这些参加语言之间的情况。也就是说，如果有 x 种语言参加分析，分析的结果是针对这 x 种语言之间的关系来说的；如果有 y 种语言参加分析，分析的结果是对这 y 种语言来说的。两种数据下得出的结论不见得完全一样。这是因为，词源统计是综合考察各种语言相互之间的距离关系，应该是一种多维空间关系，得出的树形图应该反映的是一种综合的关系。在立体空间的结构图中，各种语言之间的距离可以展现在不同的方向上；而在平面的树形图中，所有的距离都要以二维的平面形式显现，所以当增加或减少比较语言的数量时，都有可能影响其他语言之间的关系。

就目前词源统计分析方法的实践来看，这种分析方法在封闭的情况下可以揭示各种语言之间的亲缘距离和簇类，仍然不失为在语言数量相对固定的情况下分析语言之间亲缘关系的一种有效方法。

3．聚类分析与计量分类

运用词源统计分析跟统计学上的聚类分析并不相同。例如，下面是根据几种语言的距离矩阵运用 spss 软件进行的统计学分层聚类分析（Hierarchical Cluster Analysis）的结果：

图示一：



图示一显示了几种参加比较的语言（方言）之间的距离远近的类，这说明聚类分析可以在一定程度上反映不同语言之间的远近亲疏，但这种关系不同于语言的亲缘关系分析。在下面的讨论中我们会看得更清楚。

三　我国几种南亚语系语言的词源统计分析

“词源统计分析法的基本观念是两种具有亲缘关系的语言分离的时间深度，可以通过它们继承的词的共享程度来判断。”（邓、王 2003b）我们参照邓晓华、王士元对藏缅、苗瑶语族的词源统计分析的方法，主要依据 Swadesh 的 100 词表作同源词的数理分析。具体的分析原理和说明可以参见王、邓的解释说明。

1．相似矩阵（Similarity Matrix）　我们选取目前已经有一定的研究基础的几种我国南亚语系语言（方言）作为分析对象，这几种语言（方言）是：莽语、佤语、布朗语新曼俄话、布朗语关双话、德昂语、布兴语、克木语、京语、傣语，他们的材料取自相应语言的“语言简志”和“新发现语言研究丛书”后附的词汇表。

我们首先找出同源词，编制同源词表，然后计算出每对语言的同源百分比。

表一：

	莽语	佤语	新曼俄	关双	德昂	布兴	克木	京语	傣语
莽语		36	30	34	26	19	27	30	31
佤语			66	73	54	30	30	28	24
新曼俄				71	52	35	36	19	20
关双					52	39	39	24	25
德昂						41	36	20	24
布兴							45	25	20
克木								29	28
京语									32
傣 语									

说明：（1）Swadesh 的 100 核心词表我们做了如下调整：“树皮”改为“树枝”，“游泳”改为“影子”，“躺”改为“背（小孩）”，“烧”改为“烤（火）”。（2）表中所列的数字是实际出现的每对语言的对应的同源词的数量。具体的统计结果跟我们在《莽语研究》、《布兴语研究》以及其他学者的研究时的数据有些出入，主要是我们在确认同源词时采用几种语言综合比较的方法。有些词单独看时看不出有同源关系，在几种语言放在一起比较时却发现它们应该是同源词。另外，如果是合成词，只要有一个词根是同源的，就统计在内。

2．距离矩阵　由于数理树形图是通过语言间分枝的长度反映语言间的距离，所以，我们

按照 $d=-\log s$ 的公式把上面的相似矩阵转换为距离矩阵。其中 d 代表距离，s 代表相似矩阵中的数量。相似数 s 本身是一个百分数，若 100 词全部同源，即 $s=\frac{100}{100}=1$ 时，d=0，即两种比较的语言的距离为零，距离最近；当 100 词中没有一个同源时，s=0，d=∞，两者之间没有关系。也就是说，同源词的数量越大，则距离越小，两种语言分化的时间距离就越短。

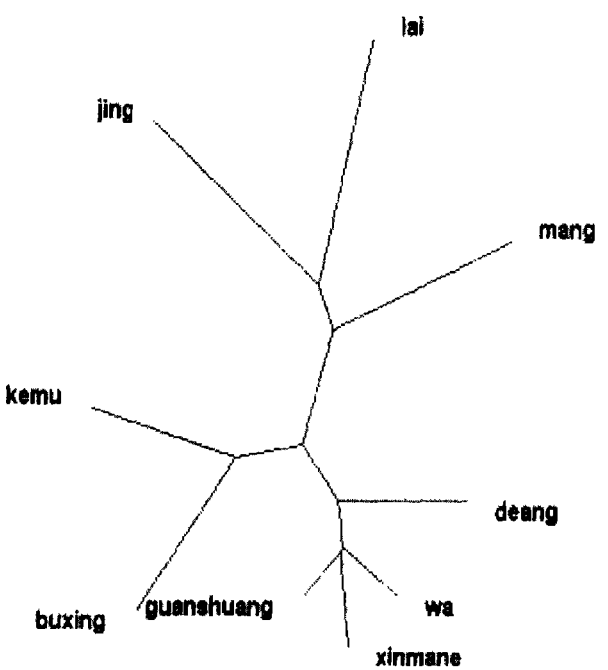
表二：

	莽语	佤语	新曼俄	关双	德昂语	布兴语	克木语	京语	侏语
莽语		0.44	0.52	0.47	0.59	0.72	0.57	0.51	0.51
佤语			0.18	0.14	0.27	0.52	0.52	0.55	0.55
新曼俄				0.15	0.28	0.46	0.44	0.72	0.70
关双					0.28	0.41	0.41	0.62	0.60
德昂语						0.39	0.44	0.70	0.62
布兴语							0.35	0.60	0.70
克木语								0.54	0.55
京语									0.49
侏语									

3．从无根树到有根树

（1）无根树　从距离矩阵转换成无根树树形图有很多种方法，前面我们用 SPSS 进行的分层聚类也是一种生成无根树的方法，但在计算反映亲缘关系的树形图中，通常采用生物学上的计算程序。我们这里使用 1987 年由 Saitou 和 Nei 发明的 Neighbor joining 程序。该程序跟其他几种类似程序一样，首先生成的是无根树。下面是用 Neighbor joining 生成的无根树。

图示二（树枝距离数值省略）：



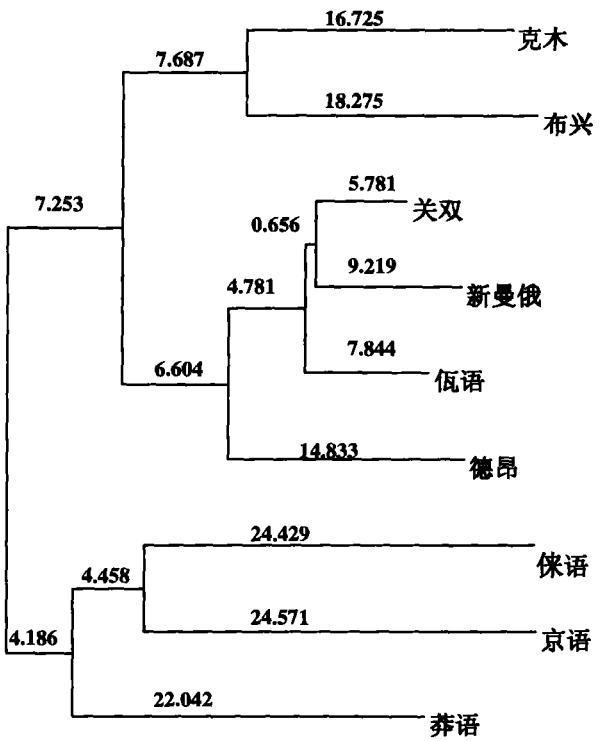
图示二是原始无根树，它建立在每对语言之间距离的基础上，大致可以反映出各种语言之间的簇属关系。我们可以以图示二为基础，选取其中分叉为一大簇或一小簇的一种或几种语言作为观察点，形成有 outgroup 的树形图。（如以莽语为 outgroup，则形成莽、京一侏、德昂一克木三大簇。）这样的树形图可以观察除 outgroup 之外的其他簇语言的关系，但看出整个树形图中所有语言的关系。为了看清楚每种语言相互之间的关系，我们必需建立一个有根树。

（2）中点生根法

邓、王（2003a：257）指出："我们可以通过在任一条分枝上插入根，使其生根的方法，来形成一棵有根的树。一般的过程是把根置于分离两个末端最远路径的中点。"我们可以把这种生根的方法叫做"中点生根法"。

不过，我们的任务是要把图示二的无根树以有根树的形式表示出来，同时又不破坏原来无根树的族簇关系。考虑到中点生根可以整体把握所要分析的各种语言，所以一般在语言亲缘关系的分析中都采用这种方法。下面是用中点生根生成的有根树。

图示三：

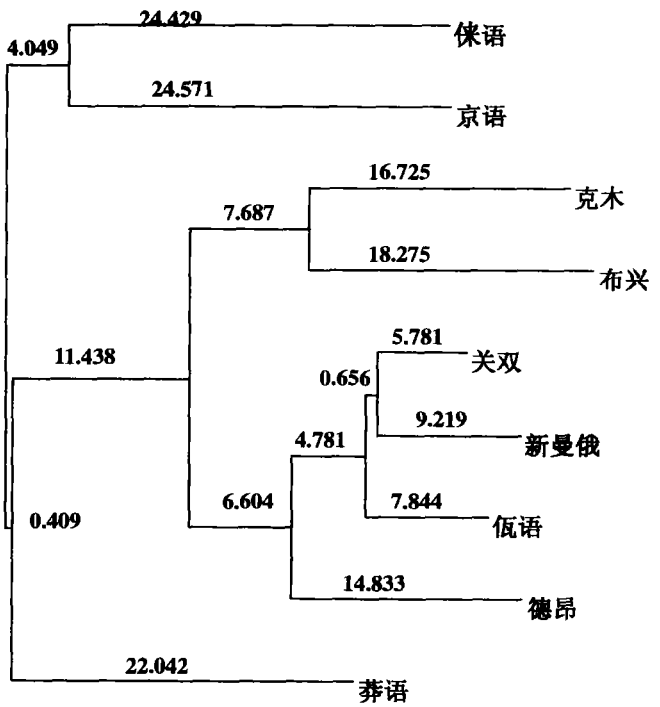


说明：（1）树形图上的数字表示距离的长短。树枝的距离只计算每一树枝的横向距离，每两种语言之间的距离等于它们之间所经过横向距离相加之和。树枝长的表示两种语言亲缘距离远，树枝短的则表示两种语言亲缘距离近。而同一个小簇类里的语言关系则比外簇类的语言关系近。（2）我们通过这种方法分析苗瑶语族语言得出的有根树跟邓、王（2003a）得出的有根树是一样的。（3）为了能用汉语形式显示各种语言，这是用 njplot 软件显示的情况。表示距离长短的数字也是该软件本身显示的。

（3）权重平均值生根法

在无根树的基础上，我们可以运用 drawgram 软件使无根树生根。该软件提供的生根方式有五种：直接分枝的中心（intermediate between their immediate descendants）、顶点权重平均值（weighted average of tip position）、最终分枝中心（centered among their ultimate descendants）、所有分枝的中心（innermost of immediate descendants）、V 字形（v-shape）。其中，顶点权重平均值按照所有族簇语言的权重平均值来确定无根树的根。其理论基础是：尽管每种语言的演化速度不同，但所有从总根分化出来的语言群演化的距离权重平均值是相同的。这是否符合语言演化的事实，还需要进一步的证明。下面是利用权重平均值生成的有根树形式：

图示四：



说明：（1）用 drawgram 画出的树形图本身不带横枝长度的数值，也不能用汉语显示各种语言的名称。我们仍用 njplot 软件显示。（2）我们在《莽语研究》中使用“语族共同词”的比较方法，即把莽语跟其他两个语族“共同”有关系的词进行比较，得出的结论是莽语跟孟高棉语族的关系要比跟越芒语族的关系近，跟权重平均值生根得出的树形图是一致的。但其合理性仍需讨论。

图示四跟图示三比较，不同处在于莽语的归属问题。按照中点生根，莽语跟京语、侏语归为一簇；按照权重平均数生根，莽语则跟克木—佤德昂归为一簇。造成这种局面的原因在于我们用来比较的语言的数量有限。如果我们增加亲属语言的数量，则会更加清除认识莽语在我国南亚语言中的具体地位。目前，我们还缺少相关其他语言的材料，无法展开这方面的比较研究。

为了给莽语的地位一个相对确切的说法，我们结合图示二显示的无根树、图示一显示的聚类分析，以及中点生根法能对局部语言综合考察的特点，倾向于图示三显示的结果，即把莽语跟京语、侏语归为一簇。这样，我们南亚语系语言就可以分为两大组：莽—京侏为一组，

克木—佤德昂为另一组。如果按照南亚语系语言的语族四分法，颜其香、周植志（1995）的分类结果跟我们运用词源统计分析的分类结果是一致的。

四　其他相关问题

关双话的地位　颜其香、周植志（1995：179－181）说："仅就《简志》中的词汇材料、语音体系来分析，关双话更接近佤语。……根据以上情况，笔者认为关双话属佤语范畴。"

在上面的树形图中也表明，关双话跟佤语的距离确实比跟新曼俄话的距离近。但在树形图中关双话却又跟新曼俄话归在了一个小簇中。[1]我们单独比较了关双、新曼俄、佤语、德昂语，其结果仍然是关双跟新曼俄归为一簇。因此我们认为，《布朗语简志》把关双话划归布朗语是有一定道理的。

五　结　语

运用词源统计分析法来对语言作数理分类，可以描述语言之间亲缘距离的远近，可以修正和补充传统语言系属分类方面的不足。但是我们同时也应看到，这种方法是建立在同源词统计的基础上的。而确定同源词的问题并没有因为这种方法的运用而得到解决。选词的范围、取舍的标准、同源词判定的方法等仍然需要多方面的传统的历史语言学知识的支持，需要每种具体语言的深入研究。不过，跟一般的数理统计相比，词源统计法借鉴了生物学上的有根树的绘制方法，使原来聚类分析显现不出来的关系能够得以明确的显现。这就为我们确定某一具体语言在相关语言中的地位和分化程度提供了明确数据，使得语言之间的相互关系更加清晰。

本文首次运用词源统计分析的方法对我国南亚语系的几种语言作分析研究，与传统分类相比较，虽然材料和方法不同，但分类的结果与传统分类大致相同。用数理方法可以显示出语言亲缘关系的相关"程度"。由于词源统计分析方法在从无根树生成到有根树时有多种不同的方法，其结果并不相同，我们分析了各种生根方法的优缺点，同时综合无根树和以往研究的成果，认为颜其香、周植志（1995）等先生所作的分类结果是可信的。同时树图研究证明把新曼俄话跟关双话划为布朗语的两种方言是有一定道理的。

参考文献

[美]G·迪弗洛思：《南亚语系》，王连清译，《民族译丛》1984年第3期。

陈国庆：《克木语研究》，民族出版社，2002年月。

陈相木、王敬骝、赖永良：《德昂语简志》，民族出版社，1986年。

[1] 不管我们用哪种生根的方法，显示的结果都是关双跟新曼俄为一簇。但当我们加进其他语言（方言）时，情况则有了变化。我们把老挝的惹茂语A方言的数据加进来统计分析的结果显示：惹茂语A方言的跟在德昂语的上面，佤语跟关双为一簇，位于最里层。这是因为相关的几种语言（方言）的数据都发生了相应的变化。不过，其他语言在树形图中的位置没有改变。

惹茂语的词汇来自林德英、史岩、谭戎·戴雅宁《惹茂（拉茂）语的两个方言》，陈相木译，《民族研究译丛》（7）第101页，云南省民族研究所编印。

邓晓华、王士元：《苗瑶语族语言亲缘关系的计量研究》，《中国语文》2003 年第 3 期。
———：《藏缅语族语言的数理分类及其形成过程的分析》，《民族语文》2003 年第 4 期。
高永奇：《莽语研究》，民族出版社，2003 年。
黄　行：《苗瑶语方言亲疏关系的计量分析》，《民族语文》1999 年第 3 期。
李道勇：《我国南亚语系诸语言纪略》，《民族研究论文集》第五集，中央民族学院民族研究所编，1985 年。
———：聂锡珍　邱锷锋：《布朗语简志》，民族出版社，1986 年。
李方桂：《中国的语言和方言》，梁敏译，《民族译丛》1980 年第 1 期。
李旭练：《仫佬语研究》，中央民族大学出版社，1999 年。
欧阳觉亚、程方、喻翠容：《京语简志》，民族出版社，1984 年。
王敬骝：《莽语调查报告》，《民族调查研究》1986 年第 4 期。
王敬骝、陈相木：《西双版纳老傣文五十六字母考释》，《民族学报》1988 年第 2 期。
颜其香、周植志：《中国孟高棉语族语言与南亚语系》，中央民族大学出版社，1995 年。
———：《佤语简志》，民族出版社，1984 年。

Abstract

This paper adopts the Lexicostatistics approach to give a numerical analysis of the genetic relationships among the Austroasiatic languages within China. According to the unrooted constructed trees and the rooted constructed trees, we confirmed the clustering of the languages and their separation in hierarchies and the genetic relationships of these languages.

（通信地址：455000　安阳　安阳师范学院中文系）

新书消息

中国少数民族语言方言研究丛书新著《玛曲藏语研究》（周毛草）、《佤语方言研究》（周植志、颜其香、陈国庆）最近由民族出版社出版；孙伯君著《金代女真语》于 2004 年 11 月由辽宁民族出版社出版；张公瑾、丁石庆主编的《文化语言学教程》于 2004 年 7 月由教育科学出版社出版；宣德五的《朝鲜语文论集》于 2004 年 7 月由开明出版社出版。

闽　于