论藏缅语亲疏关系的量化

郑玉玲

[摘要]本文旨在探讨对没有文字的民族语言，对没有或缺少历史文献的亲属语言的亲疏关系的计量分析方法。论述了藏缅语族语言亲疏关系的计量描写与印欧语系、汉语方言的不同之处，设计了适用于多种语言比较研究的语音对应规律的计量方法，提出将词汇特征统计值作为相关分析的权重处理，从而达到分类的和合理性。

ABSTRACT

the main purpose of the present paper is to approach the way of quantitative analysis for the minority languages without writings, and the degree of affinity among those genealogically related languages without or lacking historical liserature. discussing the differences of the quantitative description of the affinity among the Tibeto－Burman langusges from that among the Indo－European languages or Chinese dialects, it advances the comparative schemes for and the computational approaches to such problems in t he comparative study of the Tibeto－Burman languages as the approach to the multi－cyllabic words from the basic vocabulary, the method of quantification for sound laws, the correlative anaysis of linguistic affinity, and the statistics of the phonetic, lexical and grammatical features of the Tibeto－Burman languages each as the weight for the correlative analysis among languages. what are original in the paper are the quantitative approach to sound laws of the languages without historical literature or writings, and the analysis of the correlative matrix of sound lwas.

藏缅语(Tibeto－Burman)是汉藏语系中的一个大语族，主要分布在中国、印度、缅甸、尼泊尔、锡金、不丹、泰国、老挝等国，就国内外已记录的语言来说，超过了一百种，是汉藏语系语言的“重心”。由于藏缅语族语言内部发展极不平衡，缺乏相应的历史文献，加上许多材料是70年代以后才陆续公布，对许多材料还研究不深，因而，至今藏缅语的谱系分类在国际国内的学者中都还存在着很大的分歧。①藏缅语分类目前大多还停留在其貌相似的原始阶段，开展应用数学的方法进行系统比较，可以使藏缅语的分类科学化，使这种分类不是建筑在主观臆断上，而是建筑在可靠的定量分析上，建筑在对大量语言材料的精密分析基础上。

大家知道，语言历史比较，文献是十分重要的。印欧语系历史比较研究之所以取得成就，初去语言本身的原因，还因为有几千年连续的历史文献，如梵语、哥德语、古日耳曼语等，藏缅语众多语言中只有七世纪的藏文，十三世纪的缅文等少量古代文献，绝大多数藏缅语没有民族文字，更谈不上文献记载。对于那些无法借助书面记载的直接证据或最早的书面记载比假定的共同祖先在时间上晚的多的语言的亲疏关系研究，要借助比较法。这是美国语言学家霍凯特 (Charles Francis Hockett)的论点，②他的比较方法是比较两个语言的同源词、语法、语音和共有的词汇项目。对于我国少数民族中多数没有文字的语言来说，尤其是对藏缅语，比较语言之间词汇的语音对应规律，确定同源词所占比例，构拟同源词的原始形式，以此来判断语言间的亲疏关系是解决语言分类的重要依据。

确定真正的同源词，并不是根据语音相似，而是词的语音对应规律。例如，在藏文、拉萨藏语、夏河藏语中、“十”、“四”这两个词的语音形式：

	藏文	拉萨藏语	夏河藏语
“十”	b ʈ u	ʈ u 5 3	ʈ ə
“四”	b ʐ i	ç i 1 3	h ʐ ə

韵母主要元音的对应关系分别是/ u u ə / i i ə /，我们可以找出很多这种对应的词:

	藏文	拉萨藏语	夏河藏语
“九”	d g u	k u 3 1	h g ə
“六”	d r u g	ʈ u ʔ 1 3	ʈ ə k
“水”	ʈ h u	ʈ h u 5 3	ʈ h ə
“西”	n u b	n u ʔ 1 3	n ə p
“一”	g ʈ i g	ʈ i ʔ 5 3	h ʈ ə k
“是”	j i n	j ī 1 5	j ə n
“民族	m i r i g s	m i r i ʔ 1 3	ɳ ə r ə k
“人	m i	m i 1 3	ɳ ə

这里反复出现的韵母对应和声母对应使我们确信这些词是同源的。

针对藏缅语的特点，语言间亲疏关系的计量分析应做两方面的工作。

1、对有古文献的语言进行亲疏关系的计量分析。例如，藏语中各方言土语调查点与七世纪藏文之间以及各点间进行相似度计量。这属于语言间亲属关系的计量分析，具有历时比较研究的特点。

2、对没有共同文献的语言进行每两种语言的彼此间相似度计量，形成多语言的相似度矩阵(模糊矩阵)，经过模糊聚类法得出语言间亲疏关系量化值及树形图。这属于语言间亲疏关系的计量分析，具有共时比较研究的特点。

以上两方面的工作有大致相同的工作步骤。

1、建立藏缅语词汇语音数据库

词汇语音数据库包括国内外藏缅语主要语言和方言的基本词及其语音形式。语音有多音节标志和词素码标志，音节区分声母、韵母、声调，词的音节存储长度最长为4个音节。词的语法标志主要是词的类别，如名词、动词、形容词、量词等等，以语义范畴分类编码的方式存储。其他还有词的英义、系属标志等。这个数据库有多种用途，主要可用于藏缅语比较研究和单一语言的语音、词汇的特征统计。

2、建立语言间语音对应规律的列联表

在建立列联表之前首先要对数据库进行预处理。

通过词义识别每个语言的词根系统。对多音节词根据语义系统或形态系统找出词根或分离词根。例如“母鸡”一词在哈尼语里的语音形式是:

“母鸡”	x a 3 3	m a 3 3
	鸡	母

应分离出“鸡”和“母”两部分词根。

对多音节词根应对音节代表的语义作出标志，便于在做两个语言的多音节词比较时明确语义对应的音节。例如，藏缅语中“红花”一词，绝大多数语言中次序是“花红”，而在有的语言中可能是“红花”，这在语音比较时应避免发生错误，即使在一个语言中词的音节顺序也不一样，例如，在独龙语中构词的音节顺序正反两种情况都有:

“酸菜”	k ă n 5 5	ʈ ŭ p 5 5
	菜	酸
“铺盖”	ɟ ŏ ʔ 5 5	b ŭ ʔ 5 5
	毯子	被子
“手脚”	ŭ ɹ 5 5	x ɹ ă i 5 5

手　　脚

再有是区别词根的词缀部分。例如，在嘉绒语中，形容词一般加有前缀k ə：

“大”　k ə k t ɛ

“小”　k ə k t s i

“慢”　k ə t a l

预处理这三步工作的目的是分离出词根和确定词根在语音比较中的音节对应次序。

列联表根据有无历史文献分为两种。

第一种是有悠久历史文献的语言或已构拟出原始共同语的语言。以藏语为例，现有藏方言、土语若干点的词汇材料，可以与七世纪藏文进行基本词汇的语音对应规律的比较，形成相似条件次数表也即列联表或标定。以藏文的声母、韵母在词汇中的分布为自变量（藏文无声调），以其方言土语各点的语音形式为因变量，依次统计各方言与藏文相对应的语音形式、数量及涉及到的词项。为简便起见，下面只列出声母列联表的示意图：

		藏 文	藏方言1	藏方言2	藏方言3	.....
藏文声母	对应声母	数量　词项	数量　词项	数量　词项	数量　词项	
g	g k̇ k ṫ ʧ̇	12 0 0 0 0	0 12 0 0 0	1 2 9 0 0	0 1 8 1 2	
ʦ	ʦ	10	10	10	10	
k̇	k̇ ċ ʧ̇ c ç̇ s ṫ	16 0 0 0 0 0	11 5 0 0 0 0	8 0 5 0 2 0	7 0 6 1 1 1	
b ʧ	b ʧ ʧ	6 0	0 6	0 6	0 6	
⋮ ⋮						

表1

表中“数量”列的数值反映了各方言与藏文对应的声母在一千常用词中出现的频数。“词项”列存储着声母涉及的词项，词项以词的编码形式出现。

上述声母列联表反映了如下信息：

⑴藏文的声母与其他方言的语音对应关系。对应数量的值大，说明同源关系强，同源词多。还反应了藏方言以藏文的语音系统为参照的音类的分合与音值异同的区别。

⑵在“数量”这列中的数据反映了各方言土语与藏文语音对应的相似程度。第 1 列“数量”为藏文的声母分布情况，第 2 列“数量”是藏方言 1 与藏文声母的语音对应值，既有音值相同的数量，也有异变音值的数量。从语言学的角度，无论音值是否相同，只要语音的对应数量相对集中，或者说音变的值足够大，就反映了两方言的语音对应规律强，关系密切。所以应用表 1 我们就可以进行以藏文为自变量，以其他方言为因变量的相关分析，求出各方言与藏文的亲疏关系值。

⑶在“词项”这列中，列出了有声母对应关系的音值所涉及到的词项。例如藏文的声母“k̇ ”涉及

到的词有“细”、“缺”、“你”、“他”、“你俩”、“他俩”、..... 共 16 个词。在方言 3 中声母对应的是k' ; ʧ' ; c ç' ; s ; t' 五种, 其中变为ʧ' 的词有“缺”、“你”、“你们”等 6 个词, 我们可以说这 6 个词在声母上与藏文的对应规律较强, 因为藏文声母k' 与方言 3 声母ʧ' 有 6 个词相对应。通过韵母列联表也可以得到这种关系。把声母、韵母列联表中对应规律强的两方言共有词列出来, 就是两个方言的同源词。假如声母列联表中藏文与方言 1 对应关系强的词项有“天”、“地”、“山”、“星星”、“金子”等词项, 韵母列联表对应关系强的词项有“天”、“地”、“山”等。那么可以说“天”、“地”、“山”这三个词是方言 1 与藏文的同源词, 因为这些词的声母、韵母有对应规律。

值得注意的是, 这里所说的对应关系强即为同源词, 这种说法是一个模糊概念。强与弱的差别是在比较的过程中通过量表现出来的。

第二种列联表是对没有共同历史文献的语言, 如景颇语、藏语、哈尼语、白语、羌语等语言。把这些语言综合在一起进行比较, 这种比较具有明显的共时性, 下面我们设计这第二种列联表的计算方法, 然后再对其做进一步的讨论。

设对 n 个语言的两千常用词进行语音对应规律的比较。第一步先以语言 1 的声母、韵母、声调作为标准, 也就是做为自变量, 其余 n－1 个语言做为因变量, 分别在声、韵、调三方面同语言 1 进行比较, 形成与第一种列联表方法完全一致的一个列联表, 我们称其为 x_1 列联表。第二步以语言 2 为自变量, 其余 n－1 个语言为因变量, 与语言 2 进行语音对应关系比较, 形成 x_2 列联表, 依次进行下去, 直至完成第 n 个列联表 x_n。如示意图所示:

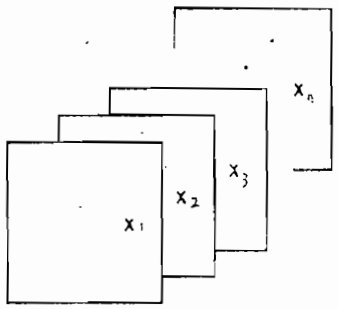

图1

上面的 n 个列联表从总体上反映了 n 个语言之间相互关系的密切程度, 因为他们之间的语音对应规律全部定量地统计出来了, 只是由于数量大, 数据多, 反映出的关系相当不明晰, 这就是我们下一个大的步骤要解决的分类问题。但可以预料到, 其统计不外乎两种结果。一是“收敛”性的状态, 一是“发散”性的状态, 我们通过大量词汇的统计可以利用模糊集合的隶属函数表征隶属程度, 从而得到比较合理的分类。

3. 相关分析

选分析两个定距变项关系的积矩相关系数 r 来测量相关的程度与变异方向。

$$r=\frac{\sum x_i y_i-\frac{1}{n}\sum x_i\sum y_i}{\sqrt{\left[\sum x_i^2-\frac{1}{n}\left(\sum x_i\right)^2\right]\left[\sum y_i^2-\frac{1}{n}\left(\sum y_i\right)^2\right]}}$$

将藏文与藏方言语音对应关系的表 1 化简为论域 $U=\{u_1u_2u_3u_4\ldots u_n\}$ 的列联表。

表中的 $u_1u_2u_3u_4\ldots u_n$ 依次是藏文、藏方言 1、藏方言 2、藏方言 3、藏方言 4... 藏方言 n。以 u_1 为 x 项,

以 u_2为 y 项，求出 $r(u_1u_2)$ 相关系数，就是藏文与藏方言 1 的相关值。再以 u_1为 x 项，以 u_3为 y 项求出 $r(u_1u_3)$ 相关系数，即藏文与藏方言 2 的相关值。直至求出 $r(u_1u_n)$，得出 n 维行向量 R：

$$R = [r_{(11)} \quad r_{(12)} \quad r_{(13)} \quad r_{(14)} \quad \cdots \quad r_{(1n)}]$$

（u1 u2 u3 u4 ... un）

简写成：$R = [r_{11}\ r_{12}\ r_{13}\ r_{14}\ \cdots\ r_{1n}]$

假如向量 R 的值是如下：

	u1	u2	u3	u4	...un
R [	1	0.8	0.6	0.7	...0.3]

每个值代表一个方言，每个值与 u_1的 1 的差值大小就是藏方言与藏文之间的亲疏程度，差值小，说明与藏文关系近，反之关系远。

对表 1 还可以进行藏方言每两个方言之间的相关系数计算，求出以藏文为参照的方言之间的亲疏关系，得出的模糊关系矩阵是如下形式：

	u1	u2	u3	u4	...	un	
	1						u1
	0.8	1					u2
	0.6		1				u3
R=	0.7			1	...		u4
	:						:
	:						:
	0.3	0.2	0.4	0.5	...	1	un

对此矩阵进行聚类分析就可以画出藏方言亲疏关系树形图。

对如图 1 所示的第二种列联表的相关系数的求法是在 n 张列联表中，用第 1 张列联表x_1求出 R_1，R 为 n 维行项量，转置得 n 维模糊列向量，即下面的模糊关系矩阵的第 1 列数据。再用第二张列联表 X_2求出第 2 个方言与其他方言的相关系数，并转置向量得出模糊关系矩阵的第 2 列数据。直至处理完第 n－1 张列联表形成模糊关系矩阵。矩阵中每个相关系数都反映了两个语言语音对应规律的强弱，它既没有以某一语言或某一历史文献的语音系统作为唯一的依据，也没有人为地划分音类，因为在没有进行大量比较研究的充分证据时是难以给不确定亲疏关系的语言归纳音类的，而这里，我们设计的多语言相关分析方法，将所有语言之间的语音对应规律量化出来，并在此基础上进行模式识别的聚类分析可以更为客观地反映方言间的实际距离。

对于有古音韵的汉语方言来说，要谈论方言的语音差别，正如语言学家吕叔湘先生所指出的，要谈的语音分歧是音类上的。③所以汉语方言亲疏关系的量化分析在汉语方言的比较研究中都是以中古音韵的音类为依据进行的比较研究，而对于藏缅语来说，我们认为语言之间，以基本词的语音对应规律的强弱（反映在相关系数上）来决定方言之间的亲疏程度是符合语言学理论的。也是多种民族语言比较研究与汉语方言比较研究在计量分析方法上的区别。

这里还要说一下，相关系数 r 是用来度量随机变量 x 和 y 之间相互关系的密切程度和方向的指标。r 值的范围在"＋1"和"－1"之间，正负号表示这两个变量之间相互关系的方向，具体到这里就是绝对值接近 1 表示两个语言密切程度高，正相关表示两个语言语音对应关系中音值相同的情况，负相关表示音值异同的情况，如果相关系数接近－1，说明两个语言关系密切，而且语音对应的音变多，但不分散，比如 x 语言的声母p 在 20 个词中变成了 y 语言声母ṕ，那么这些词在 x 和 y 语言中声母p 的相关值是－1。相关系数为 0 时表示语言间没有联系。

4、聚类分析

聚类分析的方法很多，如重心法、主成分法、类平均距离法、最长距离法、最短距离法等，由于定义类间距离的方法不同，使得分类结果不太一致。实际问题中常用几种不同的方法进行计算，比较其分类结果，选择一种比较切合实际的分类。我们这里选择最短距离分类法（也称弗洛茨瓦夫分类法）进行聚类，这种方法的特点是分类比较细，对藏缅语的初步分类，我们考虑应该选择分类细的方法，便于对语言进行谱系分类时作为参考。最短距离法的递推公式为

$$D_{h,k} = \min\{D_{h,i}, D_{h,j}\}$$

式中 $D_{h,k}$ 表示 h 类中的所有样本与 k 类中所有样本之间的最小距离，k 类是由 i 和 j 两类合并而成。

当然，最终还是要用几种不同的聚类方法进行计算，比较其分类结果，使之更适合语言的谱系分类，当然这本身又是一个研究题目。

5、特征统计

特征统计是对藏缅语的各种基本特征如复辅音、松紧元音、长短元音、辅音韵尾、声调等的数量统计及在语言中的频率分布，这应该看做是藏缅语描写研究的量化问题。这项统计的意义已不完全是共时的描写研究，而是与历史比较研究有着密切的联系。正如目前藏缅语的研究，人们对语言现状的描写，其兴趣已不仅仅是为了弄清共时的语言结构，而是为进一步探索许多共时语言的历史来源。象藏语方言的声调，现有的描写研究与历史研究相结的研究论著不断增多，如《藏语（拉萨话）声调研究》论述了拉萨话的声调从无到有、从少到多的演变过程，这个特征确立了声调的有无、多少在比较研究中的地位，④又如《藏缅语复辅音研究》一文谈到藏缅语中不少语言还保留着一定数量的复辅音，我们通过亲属语言同源词的比较，可以发现复辅音在不同的语言里变化的情况很不一样，并研究出藏缅语族各语言中复辅音演变的总趋势是简化、消失。致使目前各语言中的复辅音出现了极不平衡的状态，甚至有相当多的语言中复辅音完全绝迹。⑤这些共时与历时相结合的研究成果，在没有更多古文献作为参考的民族语言的历史比较研究中起着举足轻重的作用。也是民族语言研究较之印欧语系、汉语方言比较研究的不同之处。

特征统计的研究目的有两个，一是将特征统计的数量及在各语言中的频率分布作为前面论述过的相关分析的权重考虑，相关分析是语言之间基本词汇的语音对应强弱的反映，在进行区分方言或语言或语支之间的差别时，也就是在对语言进行不同层次的分类时，语言特征就显得很重要。我们可以将这些特征做为权重，加权后的相关分析和聚类分析将拉开语言之间不同层次的距离，使分类更为合理。第二个研究目的是作为共时描写研究的补充。因为就目前描写研究的成果来说，主要还是定性分析为主，缺乏以定量分析为依据的定性分析，这步工作不依赖计算机也是难以完成的。

本文提出了藏缅语语音对应规律的比较方法，提出了同源词的计量方法，提出了藏缅语亲疏关系的相关分析方法，并以上述方法对 38 个藏缅方言的几十个常用词进行了亲疏关系的计量分析，得出了初步结论，因篇幅关系不在此赘述。期望在藏缅语比较研究的计量分析方法研究方面与同行共同探讨。

附注

①参见戴庆厦 1990，《藏缅语族语言研究》420－425 页

参见 Paul K. Benedict, 1972, Sino－Tibeten: A Conspectus, p. 2

②参见［美］霍凯特著，索振羽、叶蜚声译《现代语言学教程》1986，207 页、10 页

③参见吕叔湘 1980，《语文常谈》89 页

④参见胡坦 1980《民族语文》1 期。藏语（拉萨话）的声调研究。

⑤参见孙宏开 1984 第十七届国际汉藏语言学会议论文，藏缅语复辅音研究。

※国家自然科学基金会资助项目