基于词汇声学距离的语言计算分类实验*

冉 启 斌

[提要] 本文使用一种新的方法来计算词汇之间的距离，即采用“动态时间规整（DTW）”算法直接计算有声词汇之间的声学距离，从而对语言进行计算分类。对8个民族语言变体和9种汉语方言进行计算分析，结果显示使用76个核心词、对应词项两两比较的方法计算效果最好。论文也尝试将得到的语言距离矩阵放入系统发生学软件进行分析，显示其操作的技术可行性。依据词汇声学距离的语言分类比以往的计算方法操作更为直接，结果更为客观。运用此方法可以进行完全自动化的语言分类。

[关键词] 动态时间规整算法　词汇距离　声学距离　语言计算分类

一　引　言

对语言进行分类可以采用不同的标准、从不同的角度进行。例如从形态类型可以将语言分为屈折语、孤立语等；从韵律类型可以将语言分为重音语言、声调语言等。在语言的诸多分类中，根据同源词对语言进行体现语言相似程度及其亲疏关系的谱系分类，是持续时间长、探讨比较深入的一种分类，原因在于谱系分类牵涉的问题复杂，涉及语言学内部共时、历时诸多方面的因素，甚至也包括语言学外部的人群来源、历史移民等因素。不管怎样，语言分类的核心思想是将相同或相近的语言变体分在一起，将不同或关系较远的语言变体分开，以体现不同语言之间的内在关系。本文介绍我们提出的一种新的语言分类方法——基于词汇声学距离的语言分类法，分析评估使用这一新的分类方法对语言进行分类的结果，并尝试探讨这种分类方法的作用和意义。

二　词汇的语音声学距离计算

语言分类方法与技术的发展，与动植物的分类方法与技术的发展具有平行性（邓晓华2006；张梦翰 2010）。以往对语言的分类大多从特征的角度进行，例如依据语言是否具有声调分为声调语言与非声调语言，依据主谓宾的语序分为SVO语言、SOV语言等。这与根据种子是否裸露分为裸子植物和被子植物、根据脚趾的数量分为偶蹄目动物和奇蹄目动物的做

* 本文为国家社科基金重大项目“中国境内语言核心词汇声学数据库及计算研究（19ZDA300）”的成果之一。黄行研究员、辛永芬教授、孙红举博士等帮助提供有关语言材料，谨致谢忱！本文曾在“《民族语文》创刊40周年学术研讨会”（北京 2019.10.12-13）报告过，感谢与会专家提出的意见！《民族语文》编辑部给出了细致精审的编校意见，在此一并致谢！

法相似，都是依据分类对象可以观察到的某些特征进行分类的。

分子生物学发展起来以后，动植物学家可以对动植物的 DNA 进行测定，得到不同物种之间的基因距离。根据基因距离数据，通过计算分析可以进行更精确、客观的动植物分类，从而在分类技术上实现了从特征到距离的发展更新。在语言研究上，依据哪些成分进行计算得到语言之间的距离数据，是语言分类方法更新的关键问题。

以往有的研究使用特定的方法将语言特征转换为距离从而进行语言计算研究（如邓晓华、王士元 2007；张梦翰、金健、潘悟云 2016）。直接对语言之间的距离进行计算，进行的比较多的是对不同语言词汇之间的距离测量。马普研究院（Max-Planck Institute）建立的“相似性自动判断程序”（Automated Similarity Judgment Program，简称 ASJP）数据库[①]依据列文斯坦编辑距离（Levenshtein Distance）计算不同语言的词汇距离（冉启斌、维希曼 2018；维希曼、冉启斌 2019），并依据词汇距离数据得到世界上 5000 多个语档（doculects）的发生学分类树形图（Müller et al. 2013），产生了广泛的影响。

ASJP 模式的编辑距离，是将不同语言的词汇形式转换为 ASJP 码，然后计算词汇的 ASJP 码之间的距离。本文中我们希望能够直接对不同语言的有声词汇（声音文件）进行声学距离计算。换言之，我们希望通过一定算法直接对不同语言词汇的录音音频文件进行距离计算，并依据这些词汇声学距离得到语言之间的距离，进而对语言进行分类。

计算声音文件之间的距离，一种重要的方法是“动态时间规整”（Dynamic Time Warping，简称 DTW）算法。它将两个声音文件看作同等长度，将声音波形转换为梅尔频率谱（Mel Frequency Spectrogram），寻找两个声音之间的最短路径，从而得到它们之间的距离（Holmes and Holmes 2001）。DTW 算法主要适用于几个音素组成的语音片段之间或单个词之间的距离计算。Mielke（2012）使用 DTW 算法计算了以英语为母语的发音人的 58 个音素（元音 17 个，辅音 41 个）之间的声学距离，并使用邻接树（Neighbor-Joining Tree）法、主成分（Principal Components）分析法绘制了 58 个音素的分类及分布情况。

本文采用 DTW 算法计算词汇之间的声学距离，不同语言之间词汇的批量声学距离计算在 Praat 6.0 中通过脚本完成。在 Praat 中 DTW 距离有两种计算方法，一种是将声音文件先转换为梅尔频率倒谱系数（Mel Frequency Cepstrum Coefficient，简称 MFCC）再进行距离计算；另一种是直接将声音文件转换为梅尔频率谱进行距离计算。经我们测试，在显示声音之间的细微差异上，后一种计算方法略好于前一种计算方法。因此，本文采用后一种计算方法。

我们对不同语言之间词汇声学距离的具体计算方式是：对 A 语言若干个词的录音文件，分别逐一计算和 B 语言若干个词的录音文件之间的声学距离。由于 DTW 算法中声音顺序对距离数值有细微影响，因此同时也计算 B 语言若干词和 A 语言若干词的声学距离。如有更多语言，则两两重复上述过程。A 语言和 B 语言最终的距离定义为两种语言各自所有词项距离的平均值；若干种语言两两之间的距离则形成距离矩阵。后期使用统计分析软件 SPSS 19.0 等对语言进行聚类等分析并作图。

在进行词汇声学距离计算之前需要对声音文件进行预处理，以使获得的声学距离更为有效。预处理的主要过程如下：（1）对声音文件进行降噪处理；（2）对声音文件进行端点检测，裁剪掉声音文件首尾空白段，提取中间的有效声音段落；（3）有的词条有 2 种以上的说法（例

[①] 该数据库为在线共享数据库，网址为：https://asjp.clld.org/。

如“星星”在有的语言或方言中既叫“星子”，也叫“星宿”），则从录音数据中提取第一种说法（往往也是最常用的说法）。对声音文件的预处理操作均使用脚本在 Praat 中自动完成。

三　不同数量词汇的声学距离计算结果比较

我们对 8 种中国境内非汉语语言变体（为方便起见下文称“语言变体”或“民族语言变体”）进行了试验。8 种语言变体分别是：哈萨克语（新疆哈巴河县），蒙古语（内蒙古正蓝旗），蒙古语（布里亚特方言，内蒙古鄂温克旗），唐汪话（甘肃东乡县），藏语（西藏拉萨市），瑶语（云南勐腊县），瑶语（海南琼中县），壮语（广西南宁武鸣区）[①]。选择这 8 种语言变体的主要原因是它们在“中国语言保护资源工程项目”（后文省称“语保项目”）中均有立项，而且我们自己或承担其中部分语言的录音工作，或能够便捷获得某些语言的录音材料。同时，“语保项目”的录音在采样率、信噪比、存储位数等方面都有统一标准，有利于使被测试的语言在相近的录音音质基础上进行声学距离计算。另外，这 8 种语言变体既有阿尔泰语系语言（突厥语族的哈萨克语，蒙古语族的蒙古语），也有汉藏语系藏缅语族语言（藏语）、苗瑶语族语言（瑶语）、壮侗语族语言（壮语），还有混合语（唐汪话），语言之间的差异较大。除此以外，这些语言变体中也有相似度很高的语言变体（例如蒙古语的两种变体，瑶语的两种变体）。这些语言变体有的亲缘距离远，有的亲缘距离近，适合于用来检测声学距离计算是否确实有效。

Mielke（2012）在计算不同音素之间的距离时使用的是 9 个样本。我们最初每种语言选择了 10 个词进行计算，即“语保项目”录音词汇“通用词表”第一部分“天文”类最前面的“太阳”“月亮”“星星”“云”“风”“台风”“闪电”“雷”“雨”“下雨”这 10 个词。Mielke（2012）计算音素之间的距离时是将所有音素两两进行比较。我们也使用了这一方法。使用这一方法还有一个依据，是 ASJP 模式语言计算的“归一化莱文斯坦距离商”（Levenshtein Distance Normalized Divided，简称 LDND），其中考虑了两两比较若干语言所有词汇项目的编辑距离这一因素，该因素可以消除由于偶然相似带来的两种语言距离接近的影响（Wichmann, Holman, et al. 2010；Holman et al. 2011）。下文我们将这一方法简称为“所有词项两两计算”。

按照上述计算方法得到 8 种语言变体的距离矩阵，使用 SPSS 19.0 进行系统聚类分析（聚类方法为组间联接，度量标准为平方欧式距离。后同不另注），结果如图 1 所示。

从图 1 可见，8 种语言变体聚类的结果并不理想，甘肃东乡的唐汪话和海南琼中的瑶语聚在了一起；正蓝旗的蒙古语和拉萨藏语聚在了一起。另外，哈萨克语、壮语的位置也不太理想。这是否说明这一方法不可行呢？Mielke（2012）计算音素之间的距离使用的是 9 个样本，而我们计算的是语言之间的距离，考虑到语言与音素情况并不相同，我们将每种语言的词汇样本增加到 30 个，亦即“语保项目”录音词汇“通用词表”的前 30 个词。仍然使用“所有词项两两计算”的方法计算得到 8 个语言变体的距离矩阵，系统聚类分析结果如图 2 所示。

[①] 为方便在统计分析软件里统一分析作图，我们按一定规则对语言变体分配了 4 个字母的代码。分配代码的规则是：采用某语言变体 ISO639-3 代码的前 2 个字母，加上该语言变体使用地区（市或县等）名称前 2 个字的首字母；没有 ISO 代码的由我们自己另行命名。这样，8 个语言变体的代码分别是：哈萨克语（哈巴河）—KAHB，蒙古语（正蓝旗）—MVZL，蒙古语（鄂温克旗布里亚特方言）—BXEW，唐汪话（东乡）—TANW，藏语（拉萨）—BOLS，瑶语（勐腊）—IUML，瑶语（琼中）—IUQZ，壮语（武鸣）—ZCWM。

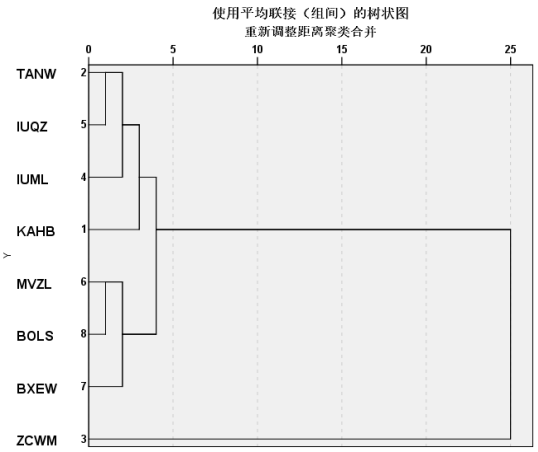

图 1　8 个民族语言变体（各 10 个词，所有词项两两计算）的系统聚类结果

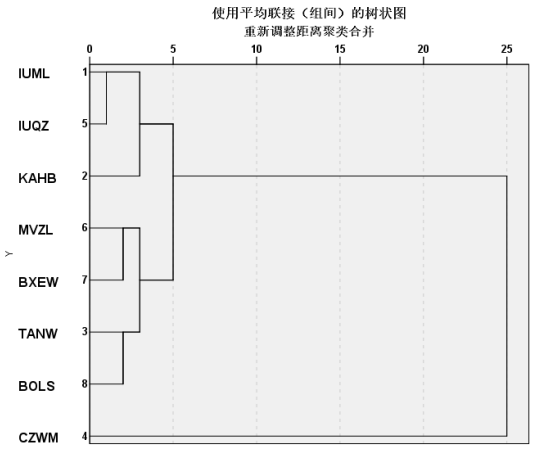

图 2　8 个民族语言变体（各 30 个词，所有词项两两计算）的系统聚类结果

从图 2 来看，好几个语言变体的聚类结果得到很大程度的改善。两种瑶语变体、两种蒙古语变体都聚在了一起。藏语、唐汪话的位置也发生了改变。当然，图 2 中也仍然存在一些不理想的地方，例如哈萨克语和两种瑶语变体的上位节点聚在一起；藏语和唐汪话的距离比较近等。

从图 1 和图 2 对比来看，增加词汇样本是改善聚类结果的有效手段，因此我们打算进一步增加用于计算的词汇项目。我们本来考虑增加到斯瓦迪士的 100 核心词，但“语保项目”的 1200 通用词表中和斯瓦迪士 100 核心词表重合的只有 76 个词，因此我们尝试增加到 76 个词汇项目进行计算。这样得到的 8 个语言变体聚类结果如图 3 所示。

图 3 中瑶语的两个变体、蒙古语的两个变体都分别聚在一起。与图 2 相比，图 3 中的主要不同是唐汪话比较接近瑶语，藏语、哈萨克语依次处在瑶语、蒙古语的外围。从总体聚类来看，整个聚类中其他语言为一大类，壮语单独为一类。在所有其他语言中，哈萨克语单独

为一类，与其他语言并列；再在其下，藏语与剩下的语言并列。这种聚类的效果怎么样，是否是这些语言变体实际情况的反映呢？

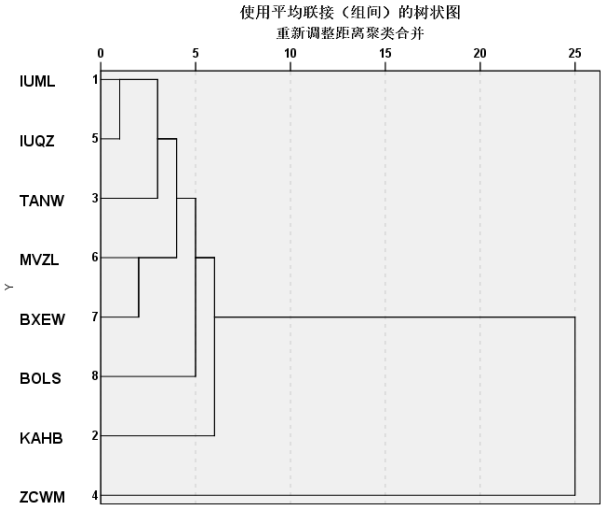

图 3　8 个民族语言变体（各 76 个词，所有词项两两计算）的系统聚类结果

四　词汇声学距离计算的两种计算方法比较

从 10 个词到 30 个词，再到 76 个词，我们推测增加词项数目的优势可能已经发挥得差不多了，继续增加样本很难有更大的改善。我们没有再增加更多的词汇项目，而是改变了“所有词项两两计算”的方法。我们改用“对应词项两两计算”的办法，即只计算 A 语言的“太阳”和 B、C 等语言相对应的“太阳”的距离，A 语言的“月亮”和 B、C 等语言相对应的“月亮”的距离（其余以此类推）。这样大大减少了计算量。仍然以 8 个语言变体的 76 个词为对象，使用“对应词项两两计算”方法得到距离矩阵，聚类分析结果如图 4 所示。

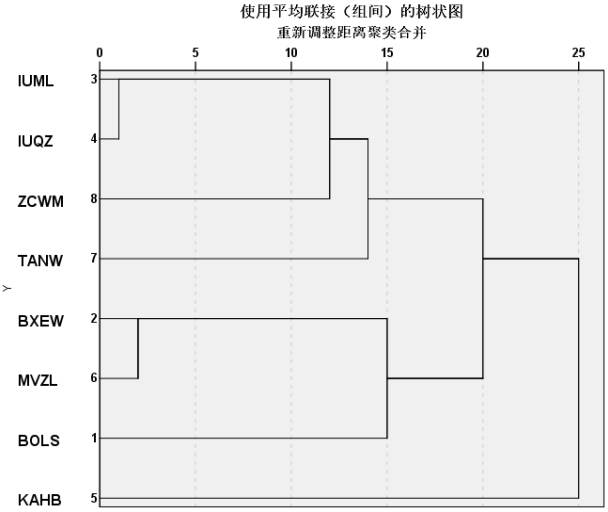

图 4　8 个民族语言变体（各 76 个词，对应词项两两计算）的系统聚类结果

图 4 显示，瑶语的两个变体、蒙古语的两个变体仍然分别聚在一起，说明“对应词项两两计算”方法没有降低距离计算效力；但是对比图 4 和图 3 可见，整个聚类结构发生了巨大的变化。在图 3 中是壮语独立为一类，其他语言聚为一类；而在图 4 中则是哈萨克语独立为一类，其他语言聚为一类。在除哈萨克语之外的语言中，瑶语、壮语、唐汪话聚为一类，蒙古语、藏语聚为一类。

对比图 3 和图 4，两种算法在聚类的整体结构上差异较大，究竟哪一种结果能更好地反映语言事实呢？单纯从这 8 个语言变体来看似乎很难作出判断。我们打算对更多语言材料进行试验。我们采用了距离比较接近的 9 种汉语方言语料进行测试，语料也来自语保项目录音。这 9 种汉语方言分别是：中原官话郑曹片河南浚县话，中原官话郑曹片河南南阳话，中原官话洛徐片河南兰考话，中原官话洛徐片河南许昌话，中原官话洛徐片河南长葛话，中原官话蔡鲁片河南漯河话，西南官话成渝片重庆合川话，西南官话成渝片重庆石柱话，西南官话成渝片重庆潼南话[①]。

每种汉语方言仍使用 76 个核心词，首先采用“所有词项两两计算”的方法，得到聚类分析结果如图 5 所示。

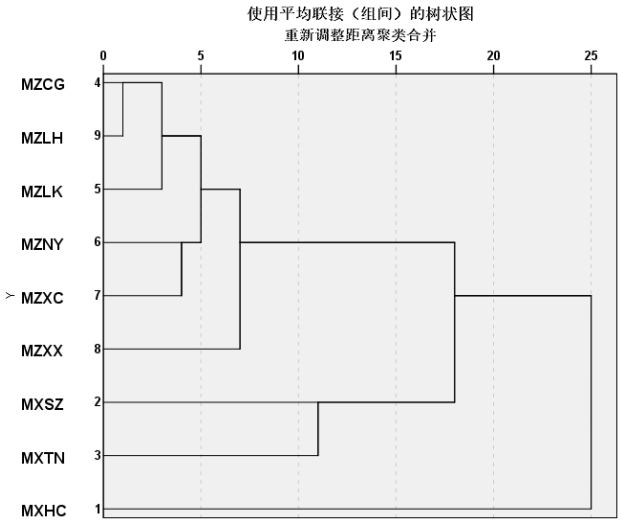

图 5　9 种汉语方言（各 76 个词，所有词项两两计算）的系统聚类结果

图 5 中聚类的总体结构分为两大类，一类为合川话；另一类为石柱话、潼南话以及其他 6 种中原官话方言，这与语言事实是存在一些差距的。在进一步的分类中，石柱话与潼南话聚在一起，浚县话与其他五种中原官话方言并列，或许有一定道理。从图上看距离最近的是长葛与漯河、南阳与许昌。南阳话与许昌话都属于郑曹片，似乎可以理解，但长葛话与漯河话则小片不同。整个聚类中与汉语方言的分类吻合程度不太高。总体来看图 5 的结果并不够理想。

① 9 种汉语方言的代码分配以 MZ 表示中原官话，MX 表示西南官话，后加县市名的首字母，分别是：中原官话郑曹片河南浚县—MZXX，中原官话郑曹片河南南阳—MZNY，中原官话洛徐片河南兰考—MZLK，中原官话洛徐片河南许昌—MZXC，中原官话洛徐片河南长葛—MZCG，中原官话蔡鲁片河南漯河—MZLH，西南官话成渝片重庆合川—MXHC，西南官话成渝片重庆石柱—MXSZ，西南官话成渝片重庆潼南—MXTN。

为进行比较，我们又使用“对应词项两两计算”的方法对 9 种汉语方言语料作聚类分析，结果如图 6 所示。

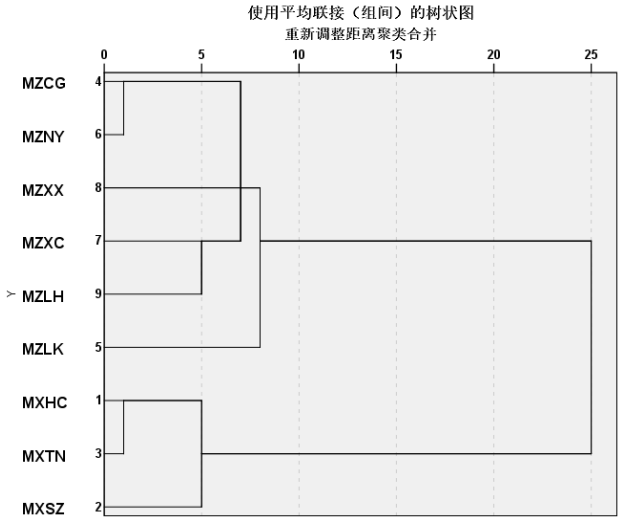

图 6　9 种汉语方言（各 76 个词，对应词项两两计算）的系统聚类结果

图 6 中 9 种汉语方言则可以分为中原官话和西南官话两大类，相比于图 5，下位层次的聚类效果也得到了一定程度的改善。不仅下位层次的聚类分散的情况减少，而且一些末端聚类也更容易解释。例如，图 6 中长葛话、南阳话聚在一起，合川话和潼南话聚在一起，二者的距离比较接近；而在图 5 中则是潼南话与石柱话聚在一起。除此以外，潼南话、合川话、石柱话虽然都属于西南官话成渝片，但潼南、合川在地理位置上临近，两地的路程不到潼南、石柱的四分之一，合川话与潼南话聚在一起更为合理。另外，3 种西南官话方言的距离总体小于 6 种中原官话的距离也比较符合实际。当然，图 6 中的一些末端聚类也有与以往小片划分不一致的地方。

从上面的情况来看，“对应词项两两计算”的方法比“所有词项两两计算”的方法效果要好。实际上，“对应词项两两计算”比较准确地反映了若干种语言变体中对应概念的声学距离；而“所有词项两两计算”则除对应概念的声学距离外，还加入了非对应概念的声学距离，这有可能影响到语言变体之间的实际距离。相较于 ASJP 模式的 LDND 距离计算方法的“所有词项两两计算”，声学距离计算中“对应词项两两计算”的方法可能更为有效。“对应词项两两计算”得到的距离相当于 ASJP 模式的“归一化莱文斯坦距离（LDN）”（参冉启斌、维希曼 2018）。

现在再来看前文的图 3 和图 4，应该认为图 4 的聚类结果更能反映 8 个民族语言变体的实际声学距离情况。从图 4 中可看到，整个聚类分为哈萨克语与其他所有语言两大类，藏语与两种蒙古语变体聚在一起，壮语与两种瑶语变体聚在一起；在 8 种语言变体中，唐汪话与壮语、瑶语的距离更近。从图 4 来看，在相对距离从 2~12 左右的范围内，8 个语言变体都可以分为哈萨克语、藏语、蒙古语、唐汪话、壮语、瑶语等 6 个类别。这 6 个类别实际上也就是 6 种语言，语言与方言的界限在图 4 中体现得非常清楚。

现在再看图 6 对汉语方言的分类情况。依据词汇声学距离的聚类，与原来的方言分区也

有不完全一致的地方，例如长葛话属于中原官话洛徐片，南阳话属于中原官话郑曹片，二者地理距离也不近，但二者的词汇声学距离却非常近。依据词汇声学距离得到的聚类结果还应该深入研究。

通过前面的试验分析可以得到一些初步的结论。“所有词项两两计算”容易造成聚类分散的情况，而“对应词项两两计算”能够更好地反映不同语言变体的距离。另外值得说明的是，“所有词项两两计算”由于需要两两比较所有词项的距离，因而计算工作量巨大；而“对应词项两两计算”则可以大幅度缩短计算时间。因此，不管从哪个角度来说，“对应词项两两计算”都优于“所有词项两两计算”。

五　依据词汇声学距离进行语言系统发生学分析

近些年，越来越多的语言学研究将分子生物学的系统发生学（phylogenetics）方法引入进来，探索语言在历史上的演化发展情况（Wichmann, Müller, et al. 2010；Holman et al. 2011；邓晓华等 2015；Zhang et al. 2019；Sargart et al. 2019）。依据词汇的声学距离可以形成语言的距离矩阵，语言的距离矩阵可以作为一些分子生物学软件的输入数据，进行语言的系统发生学分析。由于是初次研究，本文所用的语言变体数量还比较有限（8 个民族语言变体，9 种汉语方言），不过我们仍然有兴趣利用这些材料作一个初步展示，以表明依据词汇声学距离进行语言系统发生学分析的可行性。

首先使用“分子进化发生学分析”软件 MEGA 7.0 调入该软件可以识别的语言距离矩阵，使用邻接树（Neighbor-Joining Tree）法构建系统发生学树图。由于 8 个民族语言变体中有的语言之间没有亲缘关系，因此我们选用环状模式的无根树来展示 8 个语言变体的关系，如图 7 所示。

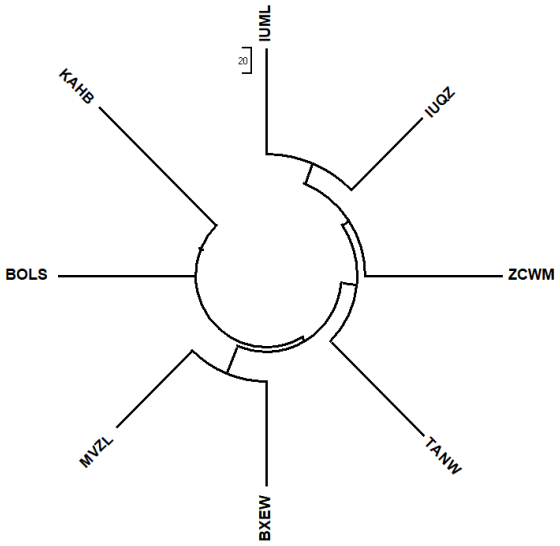

图 7　8 个民族语言变体的系统发生学树图（环状模式）

图 7 中，哈萨克语和藏语独立为 2 类；其次两个蒙古语变体聚为一类，其他语言聚为一类；再其次唐汪话独立为一类，其他语言聚为一类；壮语和两个瑶语变体处在 1 个聚类之下。图中蒙古语、瑶语各两个变体处在末端节点之下，符合语言的实际情况。唐汪话是不典型的混合语，词汇主要为汉语词汇（徐丹 2018）。唐汪话和壮语、两个瑶语变体处在一个节点之下，有可能表明汉语与壮语、瑶语的密切关系。当然，本文依据的语言变体数量比较少，更确切的分析研究还需要较大样本的语言数据支持。本文这里只在于表明依据词汇声学距离进行语言发生学分析在技术上是可行的。

系统发生学树图一般适合于展现生物种类之间的纵向遗传关系，而系统发生学网络分析则可以表现生物种类之间基因横向转移和杂交等情况。一些语言研究借助系统发生学网络来分析语言之间的相互影响和接触情况。Szeto et al.（2018）以音系、形态句法、语义、语法化模式等方面的 21 个类型学特征为基础，绘制了 42 种汉语方言的系统发生学网络图。维希曼、冉启斌（2019）依据 ASJP 模式的词汇语音形式编辑距离绘制了 65 种汉语方言的系统发生学网络图，得到的主要结论有：闽方言和吴方言相互之间以及方言内部的接触和影响都较少；北方方言和一些过渡方言相互接触和影响较多；湘、赣、客等方言过渡性特征明显等。

本文中我们使用词汇声学距离计算得到的 9 种汉语方言距离数据，在分子生物学分析软件 SplitsTree4 中进行系统发生学网络分析。使用邻网法（NeighborNet）绘制了 9 种汉语方言系统发生学网络图，如图 8 所示。

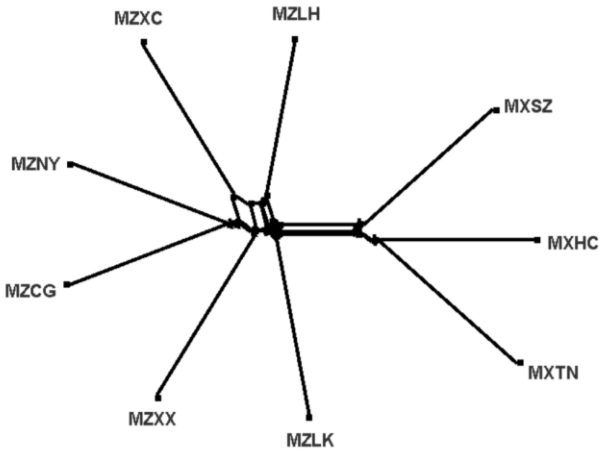

图 8　9 种汉语方言的系统发生学网络图

图 8 中左边为 6 种中原官话方言，右边为 3 种西南官话方言。图中存在一定程度的网格结构，网格结构是冲突信息的表现。所谓冲突信息是指生物种类或语言变体在演化过程中出现了与单纯纵向传递相冲突的信息，冲突信息越多表明横向传递和相互接触越多。图 8 中由于方言数量并不多，盒状网格结构不太显著。从中可以看出的是漯河话和许昌话接触相对较多，许昌话和南阳话也有一定程度的接触。更多汉语方言的接触情况可以参看维希曼、冉启斌（2019）关于 65 种汉语方言的系统发生学网络分析。这里我们主要说明依据词汇声学距离进行语言系统发生学网络分析的技术可行性。

六 结 语

一般认为语音的声学特征具有变异性大的特点，但是本文的实验显示当词汇样本达到一定数量时，依据声学距离得到的聚类分析结果与一般对语言的分类结果具有很高的吻合度。相比于以往的语言距离计算，依据词汇声学距离进行语言距离计算具有以下优点：一是距离计算更为直接，不需要将词汇的语音形式进行音标或其他编码的转写；二是距离计算更为客观，可以避免语音转写过程中采用不同标写方案带来的人为主观差异。

由词汇声学距离计算可以得到语言距离矩阵，这表明只要获得语言中一定数量的有声词汇材料，就可以实现对这些语言的完全自动分类。江荻（2017）曾按照词汇的编辑距离进行藏缅语族语言谱系的自动分类，本文使用的方法相比起来更为直接。除此以外，本文的初步分析显示，依据词汇声学距离的语言距离矩阵也可以运用于系统发生学等分析，因此这一研究方法具有多方面的应用前景。

参考文献

[1] 邓晓华. 2006.《汉藏语系的语言关系及其分类》，华中科技大学博士学位论文.

[2] 邓晓华、王士元. 2007.《壮侗语族语言的数理分类及其时间深度》，《中国语文》第 6 期.

[3] 邓晓华、杨晓霞、高天俊. 2015.《试论语言演化网络——以藏缅语为例》，《语言研究》第 3 期.

[4] 江 荻. 2017.《藏缅语谱系的自动分类实验》，《中国民族语言学报》第 1 辑，北京：商务印书馆.

[5] 冉启斌、索伦•维希曼. 2018.《怎样区分语言与方言——基于核心词汇的距离计算方法探索》，《语言战略研究》第 2 期.

[6] 维希曼，索伦、冉启斌. 2019.《ASJP 模式的汉语方言计算分析——以 65 种汉语方言语档为例》，《现代语文》第 5 期.

[7] 徐 丹. 2018.《中国境内的混合语及语言混合的机制》，《语言战略研究》第 2 期.

[8] 张梦翰. 2010.《介绍生物进化计算在汉语研究中的应用》，《东方语言学》第八辑，上海：上海教育出版社.

[9] 张梦翰、金 健、潘悟云. 2016.《闽南方言传播模式的计量分析》，《语言科学》第 5 期.

[10] Holmes, J., and W. Holmes. 2001. *Speech Synthesis and Recognition*, 2nd edition. New York: Taylor & Francis.

[11] Holman, Eric W., Cecil H. Brown, Søren Wichmann, André Müller, Viveka Velupillai, Harald Hammarström, Sebastian Sauppe, Hagen Jung, Dik Bakker, Pamela Brown, Oleg Belyaev, Matthias Urban, Robert Mailhammer, Johann-Mattis List, and Dmitry Egorov. 2011. Automated Dating of the World's Language Families Based on Lexical Similarity. *Current Anthropology*, 52(6): 841-875.

[12] Mielke, Jeff. 2012. A phonetically based metric of sound similarity. *Lingua*, 122(2): 145-163.

[13] Müller, André, Viveka Velupillai, Søren Wichmann, Cecil H. Brown, Eric W. Holman, Sebastian Sauppe, Pamela Brown, Harald Hammarström, Oleg Belyaev, Johann-Mattis List, Dik Bakker, Dmitri Egorov, Matthias Urban, Robert Mailhammer, Matthew S. Dryer, Evgenia Korovina, David Beck, Helen Geyer, Pattie Epps, Anthony Grant, and Pilar Valenzuela. 2013. ASJP World Language Trees of Lexical Similarity: Version 4. https://asjp.clld.org/static/WorldLanguageTree-004.zip.

[14] Sagart, Laurent, Guillaume Jacques, Yunfan Lai, Robin J. Ryder, Valentin Thouzeau, Simon J. Greenhill, and Johann-Mattis List. 2019. Dated language phylogenies shed light on the ancestry of Sino-Tibetan. *Proceedings of the National Academy of Sciences*, 116(21): 10317-10322.

[15] Szeto, Pui Yiu, Umberto Ansaldo, and Stephen Matthews. 2018. Typological variation across Mandarin dialects: an areal perspective with a quantitative approach. *Linguistic Typology*, 22(2): 233-275.

[16] Wichmann, Søren, André Müller, and Viveka Velupillai. 2010. Homelands of the world's language families: a quantitative approach. *Diachronica*, 27(2): 247-276.

[17] Wichmann, Søren, Eric W. Holman, Dik Bakker, and Cecil H. Brown. 2010. Evaluating linguistic distance measures. *Physica A: Statistical Mechanics and Its Applications*, 389(17): 3632-3639.

[18] Zhang, Menghan, Shi Yan, Wuyun Pan, and Li Jin. 2019. Phylogenetic evidence for Sino-Tibetan origin in northern China in the Late Neolithic. *Nature*, 569(7754): 112-115.

Language Classification: A Quantitative Approach Based on Acoustic Distances of Lexical Items

RAN Qibin

[Abstract] The Dynamic Time Warping (DTW) algorithm is used in this paper to directly calculate the distances of recorded lexical items in different languages, data obtained from which can then be used for language classification. Eight minority language variants and nine Chinese dialects in China are employed in this pilot study. It is shown in this paper that the results are the most desirable when seventy-six core words are used and the iterative pairwise comparison of lexical counterparts is respectively and thoroughly conducted for any two individual language variants among all tested language variants. This paper also conducts a preliminary feasibility test of this method by feeding the distance matrices into phylogenetics softwares for phylogenetic analysis. The results of these pilot analyses definitely indicate that, compared with previous quantitative studies on language classification, language classification based on acoustic distance matrices of lexical items is more straightforward in computing method and more objective in results obtained. The classificatory performance proves that this method is feasible in a fully automated language classification.

[Keywords] Dynamic Time Warping (DTW) algorithm lexical distance acoustic distance quantitative approaches to language classification

（通信地址：300071　天津　南开大学文学院）

【本文责编　胡鸿雁】