内部差异与外部关联

——中国民族语言 592 个语档的相似度计算分析*

冉启斌　王　帅

[提要] 本文以中国境内的 592 个民族语语档为考察对象，采用 ASJP 模式相似度计算的方法，从语言群的内部差异与语言群之间的外部关联两个方面探讨民族语的各方面表现与特点。在所考察的语档中，中国境内语言差异最大的是云南剑川金华白语和贵州锦屏偶里苗语。从语系来看，内部差异最大的是汉藏语系，最小的是印欧语系；从语族来看，内部差异最大的是印度尼西亚语族，最小的是满—通古斯语族。在语言群的外部关联方面，从语系的层面看，中国境内各个语系之间的距离都很远，其中最远的是南亚语系与印欧语系，最近的是南亚语系与南岛语系。从语族的层面看，满—通古斯语族与伊朗语族的距离最远，蒙古语族与满—通古斯语族的距离最近。如果将汉语考虑在内，则是汉语与突厥语族的距离最远，汉语与混合语的距离最近。更全面地看各个语言群之间的关系，汉语、藏缅语族、苗瑶语族的距离相对较近，壮侗语族与这三个语言群的距离较远。

[关键词] 中国民族语言　相似度　语言距离　内部差异　外部关联

一　引　言

中国自古以来就是多民族、多语言的国家，语言特征纷繁多样，语言关系错综复杂，丰富的语言资源为中国的语言研究提供了充足材料。自李方桂、赵元任建立中国语言及方言分类的基础以来，语言的分类一直普遍采用经验定性的方法（邓晓华 2006）。根据语言的发生学关系和语言共时的结构特征，国内语言学界通常将中国的语言划分为五大语系，即汉藏语系、阿尔泰语系、南亚语系、南岛语系和印欧语系，五大语系下又各自分出若干语族和语支。其中，汉藏语系和阿尔泰语系涵盖了中国境内的绝大多数语言，中国更是“汉藏语系语言分布的故乡”（孙宏开 2009）。中国语言的五大语系分类自提出以来影响至今，一直是国内中国语言谱系分类的主流观点。

作为世界上语言多样性最为丰富的国家之一，“中国语言的多样性不仅表现在语言谱系复杂，语种数量众多，并且语言（特别是汉藏语系语言）内部方言之间的相似度和可懂度也较一般意义上语言的方言变体更低”（黄行 2018）。如何区分语言与方言以及如何量化不同语言

* 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。匿名审稿专家提出了宝贵的修改意见。谨致谢忱。

变体之间的关系，是语言学研究的重要问题之一。不少研究表明，测算语言或方言之间的距离不失为一种科学有效的方法。不同语言由于成分、结构、属性等的差异从而体现出不同的亲疏远近关系，语言之间的亲疏远近关系和亲缘关系往往可以通过测量手段得到，从而形成语言的距离数据。当前，学界关于语言距离的测量维度大致可分为三类：基于结构（语音、词汇和语法等的相似性）、基于可懂度（固有可懂度和后天可懂度）和基于感知（主要为母语人的感知）。Feleke et al.（2020）深入探究了三类距离测量维度的有效性和关联性，结果发现，在语言距离的测量维度中，基于结构尤其是基于语音和词汇的语言距离测量最为科学有效。这也进一步证明了客观的语言距离测量方法需要基于语言自身的差异，在语言距离测量中应该尽量避免掺杂语言本身之外的其他因素。

本文的主要目的是以中国境内的民族语言为研究对象，使用基于核心词汇语音形式的编辑距离计算方法，从语言群内部的相似度和语言群之间的相似度，认识民族语言的差异程度及相互之间的关系。

二　研究方法与语言材料

（一）研究方法

目前语言距离计算较多采用的是通过“编辑距离”对不同语言词汇之间的距离进行测算。王璐、张吉生（2014），江荻（2017，2022），赵志靖、江荻（2018），冉启斌、索伦•维希曼（2018）等均使用编辑距离算法对中国境内的语言或方言进行过计算研究，为中国境内语言或方言的分类问题提供了新的视角。

大规模、成功运用编辑距离进行语言距离及相似度计算的是 ASJP（Automated Similarity Judgment Program，相似性自动判断程序）数据库及相关软件工具。ASJP 数据库是欧洲马普研究院（Max Planck Institute）建立的大型跨语言关联数据资源库之一，自建库以来广泛收集世界语言中斯瓦迪士核心词的语音形式，用以进行词汇语音形式相似度的计算判断，同时提供了配合该数据库的距离计算程序及代码。一种语言包含有至少 40 个核心词转写形式的文档称为一个“语档（doculect）”（原新梅等 2022）。截至 2022 年 6 月，ASJP 数据库已经发展至第 19 版（Wichmann et al. 2020），最新数据库收录有世界语言 9788 个语档材料。按照 ISO639-3 编码，这 9788 个语档涉及全球 5499 种语言（https://asjp.clld.org/）。ASJP 模式的语言距离计算现已广泛应用于语言学研究领域，如计算语言分化的起源地（Wichmann et al. 2010）、语言年代学（Holman et al. 2011）、世界语言分类（Müller et al. 2013）等。

在语言距离计算方法上，ASJP 采用“莱文斯坦编辑距离（Levenshtein Distance，LD）”的方法，计算得到的距离包括“归一化莱文斯坦距离（LD Normalized，LDN）”“归一化莱文斯坦距离商（LDN Divided，LDND）”等。相较于 LDN 距离，LDND 距离在计算较大的语档样本材料时，可以排除词长以及词汇偶然相似对距离计算产生的影响。此外，由于距离和相似度是此消彼长的关系，距离越大则相似度越小，距离越小则相似度越大，因此 ASJP 将“1-LDND%”定义为相似度指数。本文主要利用“1-LDND%”指数考察中国民族语语档之间的相似度，展现不同语言之间的差异程度（相似度数值均为百分比，为简明起见后文均省去“%”）。

（二）语言材料

本文以我们收集到的中国境内 592 个民族语语档[①]为考察对象。语档材料主要来源于中国少数民族语言简志以及少数民族语言相关的专著、期刊论文、学位论文等公开资源，另外还包括 ASJP 数据库收录的中国境内的民族语言材料等。参照《中国大百科全书》（2011）和《世界民族语言志（Ethnologue: Languages of the World）》（Simons & Fennig 2017）对中国语言谱系的分类，结合语档实际，我们将这些语档材料分入汉藏、阿尔泰、南亚、南岛及印欧等 5 个语系 9 个语族，此外还有部分语档属于混合语[②]。592 个语档的分类及数量如表 1 所示。

表 1　592 个中国境内民族语语档分类统计

语系	语族	语档数量
汉藏语系	藏缅语族	173
	苗瑶语族	148
	壮侗语族	169
阿尔泰语系	蒙古语族	28
	满—通古斯语族	16
	突厥语族	12
南亚语系	孟—高棉语族[③]	34
南岛语系	印度尼西亚语族	6
印欧语系	伊朗语族	3
其他	混合语	3

上述 592 个民族语语档均属于中国境内的民族语言，其数量不等，但我们认为这正好在一定程度上反映了中国境内民族语言数量的实际情况。如汉藏语系的 3 个语族语档数量较多，都在 150 个左右，而事实上中国境内民族语言确实是汉藏语系语言占大多数。阿尔泰语系语言的 3 个语族、南亚语系孟—高棉语族的语档数量不太多，这也与中国境内这几个语族语言的数量情况相符，与汉藏语系相比，它们的数量确实比较少。而南岛语系印度尼西亚语族、印欧语系伊朗语族的语言数量更少，这也大致与这两个语族语言的实际数量情况一致。总之，我们认为 592 个语档不平衡的分布能够较好地反映中国境内民族语言的基本面貌[④]。

三　语言群内部的相似度

语言按照其系属关系（如语系、语族、语支等）或使用的区域（主要是地理分布范围）

① 江荻（2017）对藏缅语族进行计算分类时，使用了 195 个语言或方言点的材料。

② 混合语并不属于某一个语族，为方便考虑，本文将其和民族语言 9 个语族放在一起进行分析。全文同。

③ 有的语言所属的语族存在不同的处理方式。如京语有的认为属于南亚语系越—芒语族，我们这里按照 Simons & Fennig（2017）的做法，将其归入孟—高棉语族。

④ 当然，如果要考察整个阿尔泰语系、南亚语系、南岛语系或印欧语系与中国境内汉藏语系的关系，则需要将境外阿尔泰语系、南亚语系、南岛语系或印欧语系语言都包括进来。后文在分析整个语系的关系时正是这样处理的。但本文的主要目的是分析中国境内民族语言内部和民族语言之间的关系面貌，因此在进行这部分分析时我们仍然使用 592 个语档。

会形成不同的语言群。本文主要从语言群内部的相似度和语言群之间的相似度观察中国境内民族语言的各种表现和整体面貌。我们先看语言群内部的相似度。一个语言群内部所有语档的平均相似度反映的是该语言群的总体差异程度。由于相似度与语言距离是此消彼长的关系，因此平均相似度越大，表明该语言群的总体内部差异越小；平均相似度越小，则该语言群的总体内部差异越大。一个语言群的相似度均值是该语言群内部差异状况的直观显示。下面主要从中国境内 592 个语档、不同语系、不同语族这三个层级对各语言群内部的相似度进行考察分析。

（一）民族语言的整体内部相似度

首先对 592 个语档的内部相似度进行计算考察，这是从整体角度出发考察所收集到的中国境内民族语言语档的内部相似度。使用代码程序计算 592 个民族语语档两两之间的相似度，共得到 592×591÷2=174936 对民族语之间的相似度数值。所有相似度数据的基本情况如表 2 所示。

表 2　中国民族语言 592 个语档的相似度数据概要

项目	相似度（1-LDND%）
平均值	3.80
最小值	-10.59
最大值	98.19
标准差	8.80

从表 2 来看，中国境内 592 个民族语语档之间的相似度范围分布在 -10.59 到 98.19 之间，平均值为 3.80，标准差为 8.80。我们另外计算过 300 个经过分类与地理平衡的汉语方言语档的相似度（冉启斌、丁俊 2023），在这里可以和民族语语档的数据作一个对比。汉语方言的平均相似度为 26.51，标准差为 12.16。从数据上可以看出，与汉语方言相比，民族语的相似度要低得多，且数据浮动范围也要小，表明民族语之间的差异比汉语方言要大很多，且这种较大的差异在民族语内部比较一致。

事实上，592 个语档的两两相似度数据包含了大量信息，限于篇幅，我们仅将所有相似度数据中的最大值和最小值单独拿出来进行观察。在 174936 对相似度数据中，相似度最大值为 98.19，是广西百色伶站布努语和广西百色陶化布努语之间的相似度；相似度最小值为 -10.59，是云南剑川金华白语和贵州锦屏偶里苗语之间的相似度。

广西百色伶站布努语和广西百色陶化布努语的相似度达到 98.19，几乎接近于完全相同（100）。这两个地点的布努语相似度非常高是完全可以理解的。陶化村为百色伶站瑶族乡下辖的 9 个行政村之一。按百度地图显示，从伶站瑶族乡政府所在地到陶化村仅 16.5 公里，地理距离非常近。而中国境内民族语言相似度最小值（-10.59）出现于云南剑川金华白语和贵州锦屏偶里苗语之间，这一数值已经接近世界所有语言之间的相似度最小值（-11.93）[①]。白语和苗语在语音和词汇方面差异均很大，表现出极低的相似度。

[①] ASJP 模式的相似度计算公式为“1-LDND%”，因此任意两种语言的相似度数值最大为 100（即距离为 0）；最小则可能为负值，即如果两种语言的 LDND 大于 100，会出现相似度小于 0 的情况。我们曾经以 ASJP 数据库第 19 版的早期版本为对象计算相似度，得到的世界语言 6719 个语档中的相似度最小值为 -11.93，平均值为 1.28，标准差为 4.55。这些数值可以供本文参考。

按一般的印象，中国境内语言差异最大的或许存在于阿尔泰语系语言和南亚语系语言或南岛语系语言之间，原因是它们在地理分布上距离较远。然而，数据显示事实并非如此。在我们的数据中，相似度最小值的前 40 位中都没有出现阿尔泰语系语言，直到第 43 位才是鄂伦春语与湖南江华水口镇观音山村壮语之间的相似度。这反映出中国境内民族语言中差异最大的并不是最北方的阿尔泰语系语言和最南方的南亚语系或南岛语系语言，从相似度计算结果来看，中国西南地区的语言内部呈现出了较大的差异。当然，这里我们说的是单个语言之间的差异程度，不同语系、不同语族等之间的关系是另一个问题。

（二）不同语系内部的相似度

上文以 592 个语档为对象从整体的视角观察了中国民族语言的相似度。下面从另一个层级——不同语系内部相似度的角度进行考察。我们对中国境内汉藏、阿尔泰、南亚、南岛和印欧等 5 个语系各自内部的相似度分别进行了计算，得到的结果概要如表 3 所示。

表 3　不同语系内部相似度数据概要[①]

语系	相似度均值	相似度最大值	相似度最小值	标准差
汉藏语系	4.97	98.19	-10.59	9.84
南岛语系	6.73	20.16	-1.09	6.03
南亚语系	14.50	75.33	-1.61	12.76
阿尔泰语系	15.50	94.32	-4.07	18.66
印欧语系	32.11	53.61	16.21	15.77

从表 3 可以看到，中国境内民族语言差异程度最大的是汉藏语系语言，相似度仅为 4.97，这应该与汉藏语系下属的语族、语支和语言数量有一定关系（据表 1，汉藏语系有 3 个语族，490 个语档）。汉藏语系内部的相似度最大值和最小值与表 2 所示民族语言全部语档的数值相同，这说明民族语言整体差异程度最大和最小的两对民族语均属于汉藏语系[②]。

排在第二位的是南岛语系语言，相似度均值为 6.73，其最大值和最小值分布区间比汉藏语系小得多，尤其是最大值（20.16）在 5 个语系中是最小的。特别值得注意的是，本文用以计算的南岛语系材料实际上只有属于印度尼西亚语族的 6 个语档（见表 1），这反映出一个语言群的整体差异程度并不是由用以计算的语档数量决定的，而是与这些语档实际差异程度的大小相关。南岛语系印度尼西亚语族语言在中国境内虽然分布不广，数量很少，但它们的内部差异是非常大的。

排在后面的南亚语系和阿尔泰语系语言的相似度均值（14.50 和 15.50）相差不太多，但阿尔泰语系语言的相似度分布区间比南亚语系的要大，这说明阿尔泰语系语言内部有的语言之间相似度很高，有的语言之间相似度很低，差异程度更加多样。实际上，从标准差来看，阿尔泰语系的标准差在 5 个语系中也是最大的。相似度均值最大的是印欧语系，据表 1，印欧语系实际上只有属于伊朗语族的 3 个语档，它们的总体差异程度是最低的。

总之，上述分析显示，按语系来看，中国民族语言的差异程度从高到低依次是：汉藏语

① 按相似度均值升序排列。下同。

② 本文一般所说的“汉藏语系”只包括中国境内的藏缅、苗瑶和壮侗 3 个语族，不包括汉语（汉语方言）。其他跨境语系也是如此，只包括使用地在中国境内的语言。后文全面考察各大语系的实际关系时会涵盖各个语系所包含的语言群。

系＞南岛语系＞南亚语系＞阿尔泰语系＞印欧语系。这为我们认识中国境内不同语系民族语言的内部差异状况提供了一定的新信息。

（三）不同语族内部的相似度

前文从 5 个语系的角度考察了中国民族语言的差异程度大小，下面进一步从更低的语言群层级——语族内部相似度的角度进行分析。据表 1，5 个语系的语言分属于 9 个不同的语族。此外，中国境内 592 个民族语语档还包括 3 个混合语。

原始数据太多，限于文章篇幅，我们只呈现相似度均值以及标准差数据。其中标准差有助于了解某一语言群内部相似度数据的分布情况，平均值则是反映相似度数据集中趋势的重要指标，在语言群内部可以反映各语档两两之间相似度数据的一般情况，在语言群之间则可以从总体上反映语言群与语言群之间的整体关系。不同语族相关数据的概要列表如表 4 所示。

表 4　不同语族内部相似度数据概要

语族	相似度均值	标准差
印度尼西亚语族	6.73	6.03
藏缅语族	9.87	9.90
苗瑶语族	13.27	12.64
孟—高棉语族	14.50	12.76
壮侗语族	18.06	13.97
伊朗语族	32.11	15.77
蒙古语族	34.66	16.08
混合语	35.74	10.07
突厥语族	37.10	17.96
满—通古斯语族	38.99	16.62

从表 4 可以看到，印度尼西亚语族的相似度均值最低，内部差异最大。印度尼西亚语族与表 3 中南岛语系的相似度结果完全一致。上文在考察不同语系内部相似度时，结果显示汉藏语系内部差异最大，南岛语系次之。但是汉藏语系包含 3 个语族，而南岛语系只包含印度尼西亚 1 个语族，因此按语族单独进行考察后，印度尼西亚语族的内部差异就完全显示出来了。而在汉藏语系藏缅、苗瑶、壮侗等 3 个语族中，内部差异最大的是藏缅语族，其次是苗瑶语族，内部差异最小的是壮侗语族。索伦•维希曼、冉启斌（2019）以不同的语档数量（藏缅 117 个，苗瑶 47 个，壮侗 131 个）计算了各自的相似度数据，得出的相似度均值为：藏缅（10.26）＜苗瑶（15.12）＜壮侗（23.80）。由于本文包含的语档范围与索伦•维希曼、冉启斌（2019）存在差异[①]，具体相似度均值有所不同（也存在一定的对应关系），但本文得出的 3 个语族的内部差异顺序与索伦•维希曼、冉启斌（2019）是完全一致的。换句话说，这 3 个语族无论是否包含境外语言，其内部差异的排序是相同的。

内部差异相对较小的是伊朗、蒙古、突厥、满—通古斯等 4 个语族及混合语。伊朗语族与表 3 中印欧语系的结果完全一致，这里不再解释说明。阿尔泰语系相似度在表 3 中比印欧

① 索伦•维希曼、冉启斌（2019）的语档范围包含了按当时条件能够收集到的国内外语档；本文分析的语档中，除壮侗语族有所减少外，其他两个语族则有所增加，但均只限于中国境内。

语系低，但分为 3 个语族之后，各个语族内部的相似度都比伊朗语族高。这表明阿尔泰语系 3 个语族各自的内部差异还是比较小的，其中蒙古语族的内部差异略大，突厥、满—通古斯两个语族都较小。

值得关注的是混合语的相似度数据。本文混合语 3 个语档是唐汪话、五屯话和倒话，分别分布于甘肃东乡县、青海同仁市和四川雅江县。它们虽然分布于不同的省份，然而却具有较高的相似度。这其中的原因应该在于它们都与汉语具有深厚的渊源关系，汉语成分在这 3 种混合语中都占很大的比重，从而造成它们彼此之间具有较高的相似度。

从上述分析可以看出，通过不同层级语言群内部的相似度分析，可以显示作为整体的中国境内民族语言、不同语系语言以及不同语族语言的内部差异程度，并进而展现各自的表现和特点。

四　语言群之间的相似度

语言群之间的相似度主要体现不同群之间差异程度的大小，进而显示它们之间的亲疏远近关系。我们对语言群之间相似度的考察从语系之间和语族之间两个层级进行。

（一）不同语系之间的相似度

中国境内民族语言语系之间的相似度仍然按 5 个语系进行计算，计算方法为语系 1 中的所有语档与语系 2 中的所有语档两两求得相似度，最后得到两个语系之间相似度的均值、标准差等数据。5 个语系两两之间的相似度数据量很大，相关数据概要如表 5 所示。

表 5　不同语系之间的相似度数据概要

语系 1	语系 2	相似度均值	标准差
南亚语系	印欧语系	-0.69	1.67
汉藏语系	阿尔泰语系	0.12	2.07
阿尔泰语系	南岛语系	0.14	1.91
南岛语系	印欧语系	0.22	1.76
阿尔泰语系	南亚语系	0.22	1.93
阿尔泰语系	印欧语系	0.48	2.54
汉藏语系	印欧语系	0.57	1.15
汉藏语系	南岛语系	1.38	3.02
汉藏语系	南亚语系	1.41	3.05
南岛语系	南亚语系	1.48	4.60

从表 5 可以看到，5 个语系两两之间的相似度数据整体都不太高。南亚语系和印欧语系之间的相似度均值最低，为 -0.69；最高的是南岛语系和南亚语系之间的相似度，为 1.48；其余两两语系对的相似度数据分布在二者之间。本文的印欧语系包括 3 个语档，分别是萨里库尔的 2 个塔吉克语语档和瓦罕塔吉克语语档，它们和南亚语系孟—高棉语族的 34 个语档之间的差异最大。就中国境内最大的语系汉藏语系而言，它和阿尔泰语系之间的相似度最小，为 0.12；和南亚语系之间的相似度最大，为 1.41。

（二）不同语族之间的相似度

我们仍然以 9 个语族和混合语这 10 个语言群为对象，计算不同语族之间的相似度。原始数据较多，相似度均值和标准差等各得到 10×9÷2=45 对数据。数据概要如表 6 所示。

表 6　不同语族之间的相似度数据概要

语族 1	语族 2	相似度均值	标准差
满—通古斯语族	伊朗语族	-1.06	1.82
满—通古斯语族	壮侗语族	-0.84	1.94
伊朗语族	孟—高棉语族	-0.69	1.67
满—通古斯语族	混合语	-0.39	1.42
突厥语族	孟—高棉语族	-0.36	1.73
满—通古斯语族	藏缅语族	-0.26	2.27
藏缅语族	壮侗语族	-0.26	2.23
满—通古斯语族	印度尼西亚语族	-0.18	1.98
满—通古斯语族	孟—高棉语族	-0.15	1.73
突厥语族	壮侗语族	-0.14	1.81
突厥语族	印度尼西亚语族	-0.06	1.81
突厥语族	苗瑶语族	-0.04	1.88
蒙古语族	壮侗语族	-0.02	1.91
伊朗语族	藏缅语族	0.06	2.02
突厥语族	藏缅语族	0.11	1.85
壮侗语族	混合语	0.20	2.12
伊朗语族	印度尼西亚语族	0.22	1.76
蒙古语族	藏缅语族	0.35	2.02
突厥语族	混合语	0.37	1.40
蒙古语族	印度尼西亚语族	0.42	1.87
藏缅语族	孟—高棉语族	0.47	2.27
伊朗语族	苗瑶语族	0.54	2.14
蒙古语族	苗瑶语族	0.55	2.18
藏缅语族	印度尼西亚语族	0.57	2.46
孟—高棉语族	混合语	0.60	1.95
蒙古语族	孟—高棉语族	0.68	2.00
印度尼西亚语族	混合语	0.82	1.54
蒙古语族	伊朗语族	0.83	2.37
苗瑶语族	壮侗语族	0.86	2.83
蒙古语族	突厥语族	0.88	2.31
苗瑶语族	印度尼西亚语族	1.00	2.28
满—通古斯语族	苗瑶语族	1.10	2.02

伊朗语族	壮侗语族	1.12	2.33
藏缅语族	苗瑶语族	1.33	2.63
孟—高棉语族	印度尼西亚语族	1.48	4.6
蒙古语族	混合语	1.56	2.47
突厥语族	伊朗语族	1.70	2.76
满—通古斯语族	突厥语族	1.72	1.89
伊朗语族	混合语	1.90	2.54
苗瑶语族	孟—高棉语族	1.90	2.70
壮侗语族	孟—高棉语族	1.93	3.46
壮侗语族	印度尼西亚语族	2.55	3.38
苗瑶语族	混合语	2.91	1.15
藏缅语族	混合语	4.13	4.10
蒙古语族	满—通古斯语族	6.72	2.66

表 6 的数据显示，从语族的角度来看，不同语言群之间的相似度与表 5 所显示的不同语系之间的相似度呈现出了较大差异。表 6 中相似度均值为负值的共有 13 对关系，其中最低的是满—通古斯语族与伊朗语族，为 -1.06。在相似度为正值的 32 对关系中，相似度最高的是蒙古语族与满—通古斯语族，为 6.72。

我们比较关心汉藏语系语言的情况，这里单独对和汉藏语系有关的语族情况进行观察。从藏缅语族与其他 9 个语言群的相似度来看，它与满—通古斯语族的相似度（-0.26）最小[①]，语言差异最大；与混合语的相似度（4.13）最大，语言差异最小。这应该是反映了藏缅语族与满—通古斯语族较远的距离关系，而本文涉及到的 3 种混合语与藏缅语族存在一定的关系，因此二者具有较高的相似度。从苗瑶语族来看，它与突厥语族的相似度（-0.04）最小，语言差异最大；与混合语的相似度（2.91）最大，语言差异最小。苗瑶语族与突厥语族的差异最大是可以理解的，二者分属不同的语系且地理分布上距离较远；苗瑶语族与混合语较大的相似度可能反映了二者的一些内在相关性。从壮侗语族来看，它和满—通古斯语族的相似度（-0.84）最小，语言差异最大；与印度尼西亚语族的相似度（2.55）最大，语言差异最小。壮侗语族与满—通古斯语族的差异性最大不必过多解释，而与印度尼西亚语族较大的相似性则可能反映了壮侗语族与南岛语系的某种关系。

（三）民族语言与汉语之间的相似度

行文至此，我们将引出本文另一个重要的研究参照对象——汉语。前文我们讲到，本文主要考察分析中国境内民族语言的内部差异与外部关联，因此在汉藏语系中并没有将汉语放进去。但是若不考虑汉语，对有关语系或语族之间关系的考察显然是不完备的，比如计算汉藏语系与其他语系语言的关系时，若不包括汉语则是不完善的。因此，我们将汉语和上文分析的民族语言 9 个语族以及混合语这 10 个语言群的相似度进行计算考察。

[①] 表 6 是按照相似度均值升序排列的。从表 6 看，藏缅语族和壮侗语族的相似度也是 -0.26。实际上，藏缅语族和满—通古斯语族的相似度为 -0.2580，藏缅语族和壮侗语族的相似度为 -0.2550，因本文相似度均值数据只保留两位小数，故表中呈现的结果是一样的。特此说明。

本文所用的汉语语档是经过分类与地理平衡的 300 个汉语方言语档，具有良好的代表性。使用代码程序逐一计算汉语 300 个语档与上述 10 个语言群所包含语档的相似度，其数据概要如表 7 所示。

表 7　汉语与民族语言的相似度数据概要

汉语	民族语言	相似度均值	标准差
汉语	突厥语族	-0.39	2.01
汉语	满—通古斯语族	-0.23	2.22
汉语	壮侗语族	0.26	2.91
汉语	印度尼西亚语族	0.48	2.39
汉语	蒙古语族	0.58	2.06
汉语	孟—高棉语族	0.69	2.49
汉语	伊朗语族	0.97	2.51
汉语	苗瑶语族	3.22	3.69
汉语	藏缅语族	4.69	4.44
汉语	混合语	22.48	9.31

表 7 数据显示与汉语相似度最低、差距最大的是突厥语族（-0.39）；与汉语相似度最高、差异最小的是混合语（22.48），且这一数据比其他 9 个相似度数值都高很多。前文我们曾推测，3 种混合语内部相似度较高是由于它们都具有很深厚的汉语渊源。这里得到了很好的映证，它们与汉语确实具有很高的相似度。索伦•维希曼、冉启斌（2019）曾经提出划分不同层级语言之间关系的临界值指标，其中相似度位于 18.64-50.90 之间属于相同语族之下语言的关系。按照这个指标，汉语和混合语的关系属于同一层级语言的关系。

除了混合语，汉语与藏缅语族（4.69）、苗瑶语族（3.22）都具有较高的相似度。这是容易理解的，按照通常的划分，汉语与这两个语族都属于汉藏语系。不过，汉语与壮侗语族的相似度（0.26）却很低，甚至低于汉语与伊朗语族（0.97）、孟—高棉语族（0.69）、蒙古语族（0.58）和印度尼西亚语族（0.48）的相似度。前文我们也曾提到，壮侗语族与南岛语系的印度尼西亚语族具有较高的相似度。看起来壮侗语族确实与汉藏语系的其他语族差异较大，而与南岛语系或许距离更近。

五　7 个语言群之间的相似度

我们利用更加丰富完整的语档材料来分析讨论与壮侗语族有关的一些语言群之间的关系，以汉语、藏缅语族、苗瑶语族、壮侗语族、南岛语系、南亚语系、阿尔泰语系等 7 个语言群为范围进行分析。分析的核心目的是考察汉藏语系下属的几个语族与可能相关的其他几个语系之间关系的远近，因此这 7 个语言群有的属于通常意义上的语系，有的属于通常意义上的语族。它们似乎并不在同一个层面，但我们认为为了显示这些语言群之间的实际关系，必须计算不同层面语言群的相似度才能达到考察的目的。

本文目前所用的各个语言群的语档材料均限于中国境内的民族语言。要计算各个语言群

之间真实的相似度关系，这显然是不够的。如藏缅、苗瑶、壮侗等语族语言不仅存在于中国境内，也分布于东亚其他国家（有的甚至不限于东亚）。至于南亚、南岛、阿尔泰等语系语言，更是广泛分布于中国境外的更多地区。因此对于这些语言群需要将分布在各个地区的语档都纳入其中进行计算，才能全面反映它们的实际关系。

按照这样的原则，我们从收录世界语言 9788 个语档的最新 ASJP 数据库（第 19 版）中穷尽性地将相关语言群的全部语档选入进来，并加入我们自己收集到的材料，这样得到藏缅语族 377 个语档，苗瑶语族 164 个语档，壮侗语族 232 个语档，南亚语系 182 个语档，南岛语系 1247 个语档，阿尔泰语系 144 个语档。汉语仍是经过平衡的 300 个语档。将这些语言群的语档进行相似度计算，可观察它们两两之间的距离关系。7 个语言群两两之间的相似度数据概要如表 8 所示。

表 8　7 个语言群之间的相似度数据概要

语言群 1	语言群 2	相似度均值	标准差
壮侗语族	阿尔泰语系	-0.19	1.95
壮侗语族	藏缅语族	-0.05	2.44
壮侗语族	汉语	0.05	2.75
汉语	阿尔泰语系	0.07	2.18
藏缅语族	阿尔泰语系	0.10	2.14
汉语	南亚语系	0.17	2.33
南亚语系	阿尔泰语系	0.32	2.11
阿尔泰语系	南岛语系	0.40	2.20
汉语	南岛语系	0.48	2.32
苗瑶语族	阿尔泰语系	0.51	2.10
南岛语系	藏缅语族	0.64	2.37
南亚语系	藏缅语族	0.78	2.44
壮侗语族	苗瑶语族	0.84	2.88
壮侗语族	南亚语系	0.93	2.92
苗瑶语族	南岛语系	1.01	2.39
苗瑶语族	藏缅语族	1.37	2.75
南亚语系	南岛语系	1.52	2.65
苗瑶语族	南亚语系	1.63	2.61
壮侗语族	南岛语系	1.74	2.62
汉语	苗瑶语族	3.07	3.56
汉语	藏缅语族	3.73	3.82

从表 8 数据可以看到，包含了更全面的语档之后的 7 个语言群中，相似度最低的是壮侗语族与阿尔泰语系（-0.19），也就是说 7 个语言群中真正距离最远的是壮侗语族与阿尔泰语系。其次，壮侗语族与藏缅语族的相似度也很低（-0.05）。壮侗语族和汉语的相似度（0.05）甚至略低于阿尔泰语系和汉语的相似度（0.07）。壮侗语族与苗瑶语族（0.84）、南亚语系（0.93）

的相似度略高一些。从壮侗语族的角度来看，与它相似度最高的是南岛语系（1.74）。综合来看，壮侗语族确实与藏缅语族、汉语的距离关系更远，而与南亚语系、南岛语系更近一些。

从汉语的角度来看，与其相似度最高的是藏缅语族（3.73），这是非常符合一般认知的。其次，汉语与苗瑶语族的相似度也很高（3.07），这也与一般认知相符。与此相反，汉语与南岛语系（0.48）、南亚语系（0.17）、阿尔泰语系（0.07）的相似度都较低。当然，与汉语相似度最低的是壮侗语族。

如果从藏缅语族来看，则可以看到它与其他几个语言群的相似度有高有低，最高的是与汉语（3.73），相对较高的是与苗瑶语族（1.37）。藏缅语族与南亚（0.78）、南岛语系（0.64）的相似度较低，与阿尔泰语系的相似度（0.10）更低，最低的是与壮侗语族的相似度（-0.05）。

将上述情况综合起来可以看出，汉语、藏缅语族、苗瑶语族之间都有较高的相似度，而壮侗语族与汉语、藏缅语族、苗瑶语族的距离都较远，有时甚至比这 3 个语族与阿尔泰语系之间的关系还要远。相比之下，壮侗语族与南亚语系（0.93）、尤其是与南岛语系（1.74）之间的距离要近得多。

从上述分析可以看到，无论是语系还是语族，各个语言群一方面相互之间差异程度较大，另一方面又形成关系较近的一些语言群。同时，语言群之间变化多样的相似度数据，体现了不同语言群之间错综复杂的关系。总之，我们认为相似度计算较好地展示了不同语言群之间的距离远近情况。

六 结 语

本文运用计算分析的方法，以较多数量的中国境内语言语档为材料，对不同层面民族语言的内部差异和外部关联进行了较为宏观的考察，从而对中国民族语言关系的整体表现得到了一些初步看法。如民族语言中内部差异最大的是印度尼西亚语族，最小的是满—通古斯语族；民族语言之间距离最远的是满—通古斯语族与伊朗语族，最近的是蒙古语族与满—通古斯语族；如果考虑汉语，则是汉语与突厥语族的距离最远，汉语与混合语的距离最近；等等。我们认为这些计算结果对于认识民族语言的整体表现与相互关系具有一定的参考价值，语言距离与相似度计算的结果可以与前人的相关研究结合起来，相似度计算对于发现和展示语言之间的相互关系和整体表现具有较好的作用。

本文在分析中国民族语言的内部关系与外部关联时，使用了 592 个语档材料，讨论 7 个语言群的关系时则穷尽性地使用了目前能够收集到的 2646 个语档（这个数量相当于 ASJP 第 19 版全世界语档数量的 2646÷9788≈27%）。应该说本文依据的语档数量还是比较多的，得出的结论应该具有较好的代表性。同时，由于使用的语档数量较多，我们在研究中得到了大量的计算数据，其中含有较为丰富的信息，限于篇幅等原因，这里只呈现、分析与本文讨论密切相关的极少部分数据，其他有关内容将另行报告。

一个语言群与另一个语言群之间的相似度计算与该语言群包含的下位语言群有关。如按照孙宏开等（2007:1076），壮侗语族分为壮傣、侗水、黎和仡央等 4 个语支。按本文分析显示，壮侗语族与汉语、藏缅语族、苗瑶语族的距离较远，但壮侗语族下属的 4 个语支是否都一致地与汉语等 3 个语言群距离较远，还是这些语支与汉语等的远近关系各不相同，这些问题还需要进一步的细化分析和研究。

本文使用 ASJP 模式相似度计算的方法对中国境内民族语言的差异进行考察，这些差异都是各民族语言在当代共时层面的差异。这种共时层面的差异是否能够体现各个语言系属的发生学关系，或者这种共时差异能够在多大程度上体现语言的谱系分类，则是另一个值得深入考察和讨论的问题。

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] 黄　行. 2018.《中国民族语言识别：分歧及成因》，《语言战略研究》第 2 期.

[3] 江　荻. 2017.《藏缅语谱系的自动分类实验》，载《中国民族语言学报》编委会编《中国民族语言学报》（第一辑）第 62-105 页，北京：商务印书馆.

[4] 江　荻. 2022.《汉语方言自动聚类与分区及相关计算方法》，《暨南学报》第 3 期.

[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] 冉启斌、丁　俊. 2023.《汉语方言的相似度与差异——基于 ASJP 模式语言距离计算的考察》，《语文研究》第 2 期.

[7] 孙宏开. 2009.《汉藏语系假设——中国语言学界的“歌德巴赫猜想”》，《学术探索》第 3 期.

[8] 孙宏开、胡增益、黄　行主编. 2007.《中国的语言》，北京：商务印书馆.

[9] 索伦•维希曼、冉启斌. 2019.《语言与方言的区分层级——ASJP 模式的核心词汇距离计算再分析》，《南开语言学刊》第 2 期.

[10] 王　璐、张吉生. 2014.《吴语互通度与编辑距离之间的关系》，《语言研究》第 2 期.

[11] 原新梅、丁　俊、冉启斌. 2022.《方言相似度计算与影响因素的量化——以辽宁胶辽官话为例》，《语言科学》第 4 期.

[12] 赵志靖、江　荻. 2018.《侗台语族语言的编辑距离分类》，《计算机工程与应用》第 19 期.

[13] 《中国大百科全书》编辑部. 2011.《中国大百科全书》（第二版•简明版），北京：中国大百科全书出版社.

[14] Feleke, Tekabe Legesse, Charlotte Gooskens & Stefan Rabanus. 2020. Mapping the dimensions of linguistic distance: A study on South Ethiosemitic languages. *Lingua*, 243(1): 1-31.

[15] Holman, E. W., C. H. Brown, S. Wichmann, et al. 2011. Automated dating of the world’s language families based on lexical similarity. *Current Anthropology*, 52(6): 841-875.

[16] Müller, André, Viveka Velupillai, Søren Wichmann, et al. 2013. ASJP world language trees of lexical similarity: Version 4 (October 2013). https://asjp.clld.org/download [2022-06-05].

[17] Simons, Gary F. & Charles D. Fennig (eds.). 2017. *Ethnologue: Languages of the World* (20th edition). Dallas: SIL International. http://www.ethnologue.com [2022-06-05].

[18] Wichmann, Søren, André Müller & Viveka Velupillai. 2010. Homelands of the world’s language families: A quantitative approach. *Diachronica*, 27: 247-276.

[19] Wichmann, Søren, Eric W. Holman & Cecil H. Brown (eds.). 2020. The ASJP Database (version 19). http://asjp.clld.org/ [2022-06-05].

Internal Differences and External Relevancies: The Similarity Calculation and Analysis of 592 Doculects of Ethnic Minority Languages in China

Ran Qibin and Wang Shuai

[Abstract] This paper takes 592 minority language doculects in China as the research materials, and uses the similarity calculation method of ASJP to explore the performance and characteristics of ethnic minority language groups from the two aspects of internal differences and external relevancies. Among the language doculects investigated, the languages that are most dissimilar in China are the Jinhua Bai language in Jianchuan, Yunnan, and the Ouli Miao language in Jinping, Guizhou. From the perspective of language families, the internal differences among the Sino-Tibetan languages are the biggest, while those among the Indo-European languages are the smallest. From the perspective of language branches, the internal differences among the Indonesian languages are the biggest, while those among the Manchu-Tungusic languages are the smallest. In terms of external relevancies, from the perspective of language families, various language families in China are to a great extent far from each other in linguistic distances, among which the farthest distance pairwise is between the Austroasiatic and the Indo-European languages, and the relatively closest distance pairwise is between the Austroasiatic and the Austronesian languages. From the perspective of language branches, the distance between the Manchu-Tungusic and the Indo-Iranian languages is the farthest, while that between the Mongolic and the Manchu-Tungusic languages is the closest. When the Sinitic languages are included, the distance between the Sinitic and the Turkic languages is the farthest, while that between the Sinitic and the mixed languages is the closest. From a comprehensive perspective, the distances between the Sinitic, the Tibeto-Burman, and the Miao-Yao language branches are relatively close, while the Tai-Kadai branch is far from all other three branches in terms of linguistic distance.

[Keywords] ethnic minority languages in China　similarity　linguistic distance　internal difference　external relevancy

（通信地址：300071　天津　南开大学文学院）

【本文责编　吴雅萍】