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

索伦·维希曼 冉启斌

提　要 语言与方言之间的区分具有不同的层级，本文从相似度的角度考察不同区分层级的指标。主要从 ASJP 数据库（第 18 版）分不同语系之间的语言、相同语系不同语族之下的语言、相同语族之下的语言、相同方言的不同变体四种情况选取了数量不等的语档，分别计算了各自的相似度，并进而计算了语言与方言 4 种不同区分层级的临界值，结果分别为 2.37%、18.64%、50.90%。临界值可以作为从语言内部因素划分语言与方言不同层级的参考指标。汉藏语系不同语族之下的语言其相似度总体偏低，相似度按降序为汉语方言＞壮侗语族＞苗瑶语族＞藏缅语族。

关键词 语言 方言 区分层级 相似度 临界值

一　引言

通常地说，语言是方言的上位概念，方言是语言的下位变体，两者是不同层面的概念。不过，当要确定一种语言变体是否能够独立为一种语言，还是只是作为一种方言存在，二者又是可以并称的。语言与方言的区分是十分复杂的问题，涉及到语言内部和外部方方面面的因素。本文只分析语言内部因素在区分语言与方言中的作用。

两种语言变体的差异达到怎样的程度可以认为是两种不同的语言，法兰克·布莱尔（2006）曾从语言调查的角度进行过说明。他提出，在语言调查中互相可以理解的词汇比例可以作为划分不同语言或方言的标准（后文还将说明）。冉启斌、索伦·维希曼（2018）讨论过可以以核心词汇的距离数值作为划分语言与方言的参考。

但是语言与方言之间的差异有层级的不同：（1）两种语言变体可能属于完全不同的语系；（2）两种语言变体可能属于相同语系的不同语族；（3）两种语言变体可能属于相同语族下的不同语言或语言变体；（4）两种语言变体可能只是属于相同方言的不同变体。冉启斌、索伦·维希曼（2018）只从相同语系具有不同 ISO639-3 代码的角度讨论了上述第（3）种情况。本文从语言与方言差异的层级角度对核心词汇距离计算得到的词汇相似度均值重新进行考察分析。

二　ASJP 模式的词汇相似度计算

词汇相似度有很多测量方法。王士元、沈钟伟（1992）考察相同词在不同汉语方言中的语素和构成形式，按照“双有”“有无”“双无”“无有”的方式可以计算两种方言词汇的相关系数，从而得到两种方言词汇之间的相似程度。郑伟娜（2017）在此方法基础上对语素进行了加权。杨蓓（2003）将不同汉语方言中的词分解为词段，按照声母、韵母的相似程度计算这些词汇的相关系数，从而得到不同方言的词汇相似度。这些方法立足于词汇具有千丝万缕联系的汉语方言是有效的，对于完全没有关系的语言则上述方法很难操作。

不少研究表明，依据“编辑距离”计算词汇相似度无论对有关联的语言/方言还是对没有任何关联的语言/方言都是一种有效的方法。“编辑距离（edit distance）”指由一个字符串

* 本文为北京语言资源高精尖创新中心项目“语言识别理论及语言数量统计的方法论研究”（KYR17018）子课题“基于词汇距离计算的语言分类研究”成果之一。本文通讯作者为冉启斌。**本文将刊于《南开语言学刊》2019 年第 2 期（总第 34 期）。**

转换为另一个字符串所需的编辑次数，可以用于拼写错误检查、基因相似程度的测算等。编辑距离有不同的操作方法，列文斯坦编辑距离（Levenshtein distance）对字符串只允许删除、插入和替换 3 种操作。自 Kessler（1995）使用列文斯坦编辑距离测量爱尔兰盖尔语的方言距离以来，列文斯坦编辑距离已广泛应用于语言词汇距离的计算。根据编辑距离的计算方式，得到的词汇距离准确地说是词汇的语音形式之间的距离，我们简称为词汇距离。国内王璐（2013）对 5 个吴方言点各 30 个三音节词和 20 个句子的语音形式进行了列文斯坦距离测算。江荻（2017）通过列文斯坦编辑距离考察了藏缅语族语言的谱系分类。

“相似性自动判断程序”（Automated Similarity Judgement Program，简称 ASJP）数据库（https://asjp.clld.org/）是马普研究院建立的跨语言关联数据库之一。数据库大规模收录世界范围各种语言 40 个核心词的“语档（doculect）”材料（可参冉启斌、维希曼 2017），截止目前已经发展到第 18 版（Wichmann et al2018），包括语档 7655 个，按 ISO639-3 其语言数量达到 5067 种。ASJP 数据库同时在线提供语档距离计算、系统发育分析等软件程序。

为排除词长以及词汇偶然相似的影响，ASJP 数据库提供归一化莱文斯坦距离（LDN）、归一化莱文斯坦距离商（LDND）等算法（可参冉启斌、维希曼 2017）。由于词汇的距离与词汇的相似度是此消彼长的关系，为考察不同语档之间的词汇相似度，ASJP 数据库还将“1-LDND%”定义为词汇的相似度指数。

本文主要利用“1-LDND%”指数考察不同语档之间的词汇相似度。

三　语言与方言的相似度层级

3.1 不同语系之间语言的相似度

本文开初提到，语言变体之间的差异有层级的不同。语言与语言之间差异度最大的情况存在于不同语系之间的语言。因此我们首先考察不同语系语言之间的相似度。从第 18 版 ASJP 数据库的 255 个语系中每个语系选取一个语档[①]，计算这 255 个语档的相似度。这样得到 255*254/2=32385 对语言之间的相似度，相似度数据箱形图如图 1 所示。事实上，255 个不同语系语档的相似度分布于-8.15~15.88%之间[②]，均值为 0.31%。由于这 255 个语档属于不同语系的语言，因此其中的-8.15%实际上是第 18 版 ASJP 数据库中不同语档之间相似度的最小值。除非增加新的相似度更低的语档，这个值也可以为是世界语言相似度的最小值。换言之，第 18 版 ASJP 数据库的语档相似度会分布于-8.15~100%之间（为简洁起见，后文的%非必要时均予以省略）。

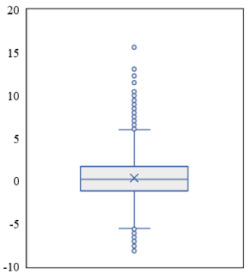

图 1　ASJP 数据库中 255 种不同语系语言之间的相似度箱形图

3.2 相同语系下不同语族语言的相似度

[①] 每个语系选取的是 ASJP 数据库中该语系下的第一种语言。如该语系下第一种语言为原始（proto-）语言，或第一种语言有效词项过少，则选取第二种语言。如第二种语言仍不符合要求，则选取第三种语言。其余以此类推。下文 3.2、3.3、3.4 选取语言的原则与此相同，不另注。

[②] ASJP 模式的相似度是按编辑距离计算得到的，两种语言之间的差异过大则编辑距离 LDND 可能大于 100，这样相似度（1-LDND%）则可能出现负值。按照编辑距离计算方法，两种语言如果完全相同则 LDND 为 0，此时相似度达到最大，为 100%。

比不同语系下语言之间相似度大的是相同语系下不同语族语言之间的相似度。为考察不同语族语言之间的相似度，在 ASJP 数据库中的 10 个语系下，每个语族选取一个语档，计算各语系不同语族语档的相似度均值。10 个语系的情况、语档数量及相似度均值如表 1 所示。从表 1 来看不同语系的相似度存在差异，数值在 3.03~12.69 之间。尼日尔-刚果、闪含、阿尔泰等语系下不同语族语言之间的相似度比较小；达罗毗荼、乌拉尔、南岛等语系下不同语族语言之间的相似度比较大；其余语系居中。10 个语系的相似度均值为 7.9，相比于 3.1 看到的不同语系之间语档的相似度（0.31）上升了 7.59。从均值上看不同语系之间的语言和相同语系不同语族之间的语言，其相似度存在一定的区分度。

表 1　相同语系下不同语族语档的相似度

语系	相似度值 1-LDND%均值	样本数量
尼日尔-刚果语系	3.03	22
阿尔泰语系	3.60	3
闪含语系	3.94	12
印欧语系	5.68	8
Sko 语系	7.29	4
南亚语系	9.32	12
Gunwinyguan 语系	9.67	7
南岛语系	11.69	34
乌拉尔语系	12.12	6
达罗毗荼语系	12.69	3
均值	7.90	11

3.3 相同语族下语言的相似度

冉启斌、维希曼（2018）给出了语言与方言区分的临界 LDN 距离值。由于 LDN 距离不够直观，本文我们从相似度值进行考察。对语言与方言的判定是十分复杂的问题。冉启斌、维希曼（2018）以是否具有相同的 ISO 码作为区分语言与方言的前提，这在很大程度上仍然是一种人为的标准。本文拟绕开这一问题，直接计算相同语族下所有语档的相似度，以获得相同语族下语言相似度的总体印象。我们从第 18 版 ASJP 数据库中重新选取 10 个语族下的所有语档，计算得到的结果如表 2 所示。

表 2 显示，相同语族下语档的相似度分布在 22.95-53.77 之间，其中 Bosavi、North_Halmahera 等语族相似度比较低；Ijoid、Turkic 等语族相似度比较高。相同语族之下语言的相似度均值为 39.07，比 3.2 看到的不同语族语言的相似度（7.90）高出了 31.17。这主要是因为将相同语族下的所有语档都计算进来了，其中包括一些相似度很高的语言变体（但并不包括相似度非常高的相同方言变体，如下文 3.4 所示的情况）。如果将这一因素考虑进去，相同语族下语档的相似度会有所降低[①]。相同语族下所有语档的相似度反映的是该语族语言的总体相似度，能给人该语族语言差异度大小的直观印象。

表 2　相同语族下不同语言语档的相似度

语系	相似度值 1-LDND%	样本数量

[①] 值得补充说明的是，冉启斌、维希曼（2018）计算了同语族之内具有不同 ISO 代码和具有相同 ISO 代码语档的 LDN 距离临界值，结果为 0.48，我们重新计算了这些语档的相似度（1-LDND%）临界值，结果为 34.7。

	均值	
Bosavi	22.95	13
Eleman	34.05	19
North_Halmahera	34.37	16
Eastern_Trans-Fly	34.52	18
Mayan	34.76	106
Huitotoan	40.77	10
Kadugli-Krongo	43.72	11
Japonic	44.07	31
Turkic	47.72	56
Ijoid	53.77	34
均值	39.07	31.4

值得说明的是本文的相似度数值与有的研究提到的相似度并不相同。法兰克·布莱尔（2006）曾提出：“如果词汇比较的结果显示，两种言语变体之间的相似程度低于 60%，那么，这两种言语变体就可视为不同语言。”布莱尔是讲使用词表进行语言调查时提到上述数值的，这一词汇相似度表示的是两种语言变体使用相同词汇的比例。而本文所说的词汇相似度是基于词的语音形式进行编辑距离计算并进行转换的结果。布莱尔所说两种语言变体的相同词汇主要是一种人为判断。如果实在要进行比较，两种语言词汇相似度低于 60%这一标准是很宽松的。按照这一标准，将会有非常多的方言变体被判断为独立的语言。

3.4 相同方言不同变体的相似度

在 ASJP 数据库中还收录有一定数量完全相同的方言的变体。这些语档大多属于同一个方言，具有相同的经纬度和人口数据等，只是核心词的来源和转写存在差异。在 ASJP 数据库中它们具有相同的语言名称，在语言名称之后用“_1”“_2”进行区分。毫无疑问，这些语档之间具有最小的相似度。我们从数据库中选取了 20 对带有“_1”“_2”的语档进行相似度计算，得到的结果如表 3 所示。表 3 看到，这些相同方言的相似度数值分布范围很大（28.99~98.72），跨越度达到 70 左右。其中 Paez、Huave 等语言变体的相似度最小；Heyo、Malagasy_Sakalava 等语言变体的相似度非常大，Heyo 语的两个语档更是接近于完全相等了。所有语言变体的相似度均值是 62.38，这个数值可以作为相同方言不同变体的参考指标。

表 3　相同语族下不同语言语档的相似度

语档 1	语档 2	相似度
Paez_1	Paez_2	28.99
Huave_1	Huave_2	31.17
Nama_1	Nama_2	35.04
Mangbetu_1	Mangbetu_2	48.2
Nihali_1	Nihali_2	48.98
Kusunda_1	Kusunda_2	49.69
Abau_1	Abau_2	50.82
Yimas_1	Yimas_2	53.83
Ama_1	Ama_2	56.28
Ik_1	Ik_2	57.73

Sunwar_1	Sunwar_2	62.89
Murrinh_Patha_1	Murrinh_Patha_2	65.74
Kaera_1	Kaera_2	67.26
Ambele_2	Ambele_1	67.76
Awngi_1	Awngi_2	80.84
Parji_1	Parji_2	82.55
Mailu_1	Mailu_2	83.15
Purhepecha_1	Purhepecha_2	87.7
Malagasy_Sakalava_1	Malagasy_Sakalava_2	90.18
Heyo_1	Heyo_2	98.72
均值		62.38

四　汉藏语系语言的词汇相似度

汉藏语系是继印欧语系之后的世界第二大语系，汉藏语系语言包含的范围有不同的意见。从李方桂（1937）以来的传统看法认为汉藏语系包括汉语族、藏缅语族、苗瑶语族和壮侗语族（如罗常培、傅懋勣 1954，马学良主编 2003 等）。有的学者认为苗瑶语族、壮侗语族不属于汉藏语系（如 Benedict1942、Matisoff1997 等）。最近发表的 Zhang et al（2019）、Sagart et al（2019）都没有将苗瑶语族、壮侗语族包括进去。作为依据距离计算的词汇相似度分析，我们仍然将苗瑶语族、壮侗语族放进了进行考察。

从 ASJP 数据库中选取上述 4 个不同语族的语档，计算各个语族的平均相似度数值。其中汉语族（汉语方言）在 ASJP 数据库中语档数量较少（包括中古汉语、上古汉语、东干语在内一共只有 19 个语档），是另行使用我们自己建立的 65 个汉语方言语档（索伦 •维希曼、冉启斌 2019）进行计算的。这样得到的汉藏语系 4 个语族语档的相似度结果如表 4 所示。

表 4　汉藏语系不同语族语档的相似度

语族	语档数量	相似度均值	分布范围
汉语族（汉语方言）	65	28.98	1.33-73.59
藏缅语族	117	10.26	-4.95 - 94.26
苗瑶语族	47	15.12	-3.7 – 91.67
壮侗语族	131	23.80	-5.32 – 89.29
均值	90	19.54	--

表 4 显示，汉藏语系语言相似度总体来看都不高。对比前文 3.3 得到的相同语族下语档的相似度，汉藏语系语档的相似度在数值分布范围、均值方面显然都要低不少。尤其是均值方面，汉藏语系语档平均（19.54）比 3.3 得到的其他语族的均值（39.07）低近 20。汉语方言（28.98）在汉藏语系中还是相似度比较高的，也比其他语族的均值低 10.09。壮侗语族语档相似度只有 23.80（低 15.27）；苗瑶语族语档只有 15.12（低 23.95）；藏缅语族语档最低，只有 10.26（低 28.81）。藏缅语族语档的相似度甚至接近于 3.2 计算的相同语系下不同语族语言的平均相似度，只比相同语系下不同语族语言的相似度均值（7.90）略高了 2.36。

相同语族下语档的相似度均值反映的是该语族的总体差异情况。不同语系、不同语族的情况很不一样，相似度数值差异很大，不能一概而论。汉藏语系语言中藏缅语族语言的差异度是最大的，汉语方言的差异度相对而言小一些。如果认为苗瑶语族和壮侗语族属于汉藏语系，则它们的差异度处于藏缅语族和汉语方言之间。单纯从相似度数据很难看出苗瑶、壮侗语族与汉语、藏缅语族之间有什么差异。

另外，不同语族语言的相似度存在差异也许与这些语言所处的地形地貌有关。藏缅语族语言所使用的区域高山大川较多，往往交通不便，人群交际频率与密度相对低一些，这或许是造成其相似度较低的原因。与此相反，汉语方言相对而言使用在更多的平原地区，交通与传播相对便利，方言之间交流更加频繁，从而使得不同方言之间相似度得到增加。杨露、余金枝（2016）从云南省玉龙县九河乡普米语的生态保护来看地理环境对语言功能演变的影响，认为崎岖的山地、常年寒冷的冰天雪地、干旱少雨的沙漠地区、原始森林、大片沼泽地区等地理区域由于自然条件恶劣、交通不便, 使得语言相对更封闭, 更有利于弱势语言的保存。从这个角度来看自然环境是会对某地区语言的相似度产生影响的。当然，语言相似度与地形地貌之间的关系还需要大量语言事实与数据的研究证实。

五　语言与方言不同区分层级之间的临界值

上文按照不同语系之间的语言、相同语系不同语族之下的语言、相同语族之下的语言、相同方言的不同变体等四种情况计算了语言与方言不同区分层级的相似度平均值。这些相似度数据可以作为认识语言与方言不同层级差异的一个参考指标。我们更感兴趣的是，如果能够给出不同语系之间的语言、相同语系不同语族之下的语言、相同语族之下的语言、相同方言的不同变体之间四者之间的临界值，对于判断语言与方言的不同区分层级可能会更有意义。如果能够给出 3 个相似度临界值，形成 4 个数值区间，按照两种语言的相似度所处的区间可以判断它们可能属于四个层级中哪一个层级的区分。

按照这一想法，我们重新计算了 4 个层级的相似度数据分布范围。某些层级的相似度数据分布范围跨度较大，我们按照四分位区间距离（1/4-3/4）给出数据的范围（这也是 Excel 默认给出的箱形图的范围）。这样得到的结果如表 5 所示。

表 5　语言与方言区分 4 个层级的相似度范围

不同层级的语档区分	不同语系之间的语档	相同语系不同语族之间的语档	相同语系相同语族之间的语档	相同方言的变体
相似度范围	-1.19~1.70	3.03 ~ 12.95	24.32 ~ 52.64	49.16 ~82.12

从表 5 可以看到，有的层级之间数据不是连续而是断裂的，如（-1.19~1.70）和（3.03~12.95），其间一端的 1.70 和另一端的 3.03 之间还存在差距；而有的层级之间数据是相互交叉重叠的，如（24.32～52.64）和（49.16～82.12），其间一端的 52.64 和另一端的 49.12 存在重合部分。为求 4 个不同层级之间的临界值，我们采用相邻层级端点数据求平均值的办法。这样得到 4 个层级之间的 3 个临界值分别为：(1.70+3.03)/2=2.37；(12.95+24.32)/2=18.64；(52.64+49.16)/2=50.90。这样，如果两种语言的相似度小于 2.37，这两种语言可以认为属于不同语系的语言；两种语言的相似度大于 2.37 小于 18.64，这两种语言可以认为属于相同语系不同语族的语言；两种语言的相似度大于 18.64 小于 50.90，这两种语言可以认为属于相同语系相同语族的语言；两种语言的相似度大于 50.90，这两种语言可以认为属于相同方言的变体。语言与方言的不同区分层级对应于不同的相似度区间可以直观地表示为图 1：

图 1 的 3 个临界值对于区分语言与方言的不同层级仍然只是一个指导性指标，不能一概而论。语言/方言的区分，以及它们的不同区分层级，是一个包含语言内部和语言外部多

方面因素的问题，在语言外部涉及政治、宗教、历史、社会等多方面的因素。因此我们得到的临界值只是从语言相似度这一内部因素出发的参考性指标，可以在不涉及其他因素时为语言/方言的层级区分提供参考。

六　结语

本文对语言与方言之间的界限做了更细的划分，将语言与方言之间的关系分为不同语系之间的语言、相同语系不同语族之下的语言、相同语族之下的语言、相同方言的不同变体等四种情况进行考察，给出了各自的相似度数据。论文并通过一定的算法计算出四种情况的 3 个临界值，临界值更便于直观地划分语言与方言的区分层级，可以为语言与方言的层级划分提供语言内部因素指标。在语系、语族之下一般还有语支，语支之下是语言，但是语支的情况十分复杂，一时难以计算其相似度数据及临界值，只能俟诸将来。语言与方言的不同层级区分在不同语系的语言中情况都可能不一样，本文得到的临界值只是一个普遍的参考指标。划分语言与方言的区分层级仍然需要结合不同语言及其使用者各自的社会、历史、政治、种族等多方面的因素决定。

参考文献

法兰克・布莱尔（2006）《双语调查精义》，卢岱译，北京：民族出版社。

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 期。

罗常培、傅懋勣（1954）《国内少数民族语言文字的概况》，《中国语文》3 月号。

马学良主编（2003）《汉藏语概论》，北京：民族出版社。

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，50-58 页。

索　伦・维希曼、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 6 期。

王士元、沈钟伟（1992）《方言关系的计量表述》，《中国语文》第 2 期。

杨　蓓（2003）《吴语五地词汇相关度的计量研究》，《语言文字应用》第 1 期，120-130 页。

杨　露、余金枝，2016，地理环境对语言功能演变的影响——以九河乡普米语小语种的生态保护为例，《青海民族研究》第 4 期，197-200 页。

王　璐（2013）《语言距离与吴语互通度》，华东师范大学外语学院博士学位论文。

郑伟娜（2017）《四邑方言词汇相似度比较分析》，《中国语文》第 6 期，693-703 页。

Benedict, Paul K. (白保罗 1942) Thai, Kadai and Indonesian: a new alignment in south east Asia, *American Anthropologist*, Vol. 44, No. 4, 576-601.

Kessler, B.（1995）Computational dialectology in Irish Gaelic. Proceedings of the 7th conference of European chapter of the Association for Computational Linguistics, 60-66, Dublin: Morgan Kaufmann.

Li, Fang-kui（李方桂）(1937) Languages and Dialects, *Chinese Year Book*, 59-65.

Mtaisoff, James A. (1997) Issues in the subgrouping of Tibeto-Burman in the post-Benedict Era, Paper on 30th ICSTL.

Sagart, Laurent, Guillaume Jacques, Yunfan Lai, Robin J. Ryder, Valentin Thouzeau, Simon J. Greenhill & Johann-Mattis List (2019) Dated language phylogenies shed light on the ancestry of Sino-Tibetan, PNAS,

Zhang, Menghan, Shi Yan, Wuyun Pan & Li Jin, Phylogenetic evidence for Sino-Tibetan origin in northern China in the Late Neolithic, Nature,

Hierarchical discriminations between languages and dialects:

ASJP approach for similarity calculation

Søren Wichmann; Qibin Ran