China Family Panel Studies

CFPS

中国家庭追踪调查

技术报告系列：CFPS-28

系列编辑：谢宇　责任编辑：张聪

中国家庭追踪调查

方言编码

武玲蔚 张文佳

2014.11.15

一、 CFPS 方言编码的内容

CFPS2010 年中记录方言使用的问题包括以下几个来源：

（1） 成人问卷

题号 D2 （变量名 QD2：您平常与家人交谈主要使用什么语言）

（2） 少儿问卷

题号 K2 （变量名 WK2：您平常与家人交谈主要使用什么语言）

（3） 公共模块

题号 S3 （变量名 KS3：您在学校与同学日常交谈主要使用什么语言）

CFPS2012 年中记录方言使用的问题包括以下几个来源：

（1） 成人问卷

题号 Z103 （变量名 QZ103：访问过程中主要使用哪种语言）

题号 Z104 （变量名 QZ104：具体是什么方言）

（2） 少儿问卷

a） 父母代答部分

题号 Z103 （变量名 KZ103：访问过程中主要使用哪种语言）

题号 Z104 （变量名 KZ104：具体是什么方言）

b） 10-15 岁少儿自答部分

题号 Z103 （变量名 KZ103_B_2：少儿访问过程中主要使用以下哪种语言）

题号 Z104 （变量名 KZ104_B_2：具体是什么语言）

二、 CFPS 方言编码的依据

1. 《中国语言地图集》

CFPS 方言编码的主要依据是《中国语言地图集》（The Language Atlas of China，以下简称《地图集》）。《地图集》由中国社会科学院语言研究所和澳洲人文科学院合作，由中国社会科学院语言研究所李荣、熊正辉、张振兴担任主编，于 1983 年开始编制，1987 年完成。《地图集》在全面的语言学调查的基础上，按古入声字、古浊声母字的演变规律对汉语方言进行分类，相比其他分类方法更为科学，已成为方言学界实际上的学科标准。《地图集》有中文和英文版本，中文版由香港朗文（远东）出版公司于 1987 年和 1991 年分两次出版。

从内容上讲，《地图集》包含了 35 副四开彩色地图，每幅图可以分为三个部分：A 组图提供了中国境内汉语方言和少数民族语言的地区分布，B 组图提供了汉语方言的分布情况，C 组图提供了少数民族语言的分布情况。每一幅图均配有详细说明。

2. William Lavely（2012）的工作

华盛顿大学的 William Lavely（2012）首次将《地图集》的方言分布信息按照 1990 年中国人口普查的县级市编码（GB 2260-88）进行整理。虽然《地图集》包含了中国的五大语系：汉藏语系（其中包括汉语和藏缅语族）、阿尔泰语系、南亚语系、南岛语系、印欧语系，但 Lavely（2012）仅对汉藏语系中的汉语语族进行了编码[1]，其过程具体如下。首先，汉语语族可以被分为十个大类：

a） 官话大区

b） 晋语区

c） 吴语区

d） 赣语区

e） 湘语区

f） 闽语大区

g） 粤语区

h） 客家话区

i） 徽语区

j） 其他话区

这一划分基本遵循了《地图集》的划分，但是也存在以下区别：其一是将官话和闽语都列为“大区”；其二是将平话、儋州话、乡话、韶关土话等四个组统一归为其他话区（因为这些组的规模都比较小，一般仅仅在一两个县级市中使用）。

然后，语言编码由六位数字组成，其中各个数位的分布为：

语系(1 位)

语族(1 位)

大区 (Supergroup or Group)（1 位）

区片（Group or Subgroup）(1 位)

[1]即使如此，其编码可以根据《地图集》对非汉语语言进行扩展。

片（Subgroup）(2 位)

由于这里的编码主要考虑到汉语语族的语言分布，因此六位编码的前两位总是 11（汉藏语系中的汉语语族），从而前两位可以被省去。因此，Lavely（2012）在数据中的语言编码是四位数：第一位代表大区或者区（官话大区，晋语区、吴语区，等等）；第二位代表区或者片（比如，东北官话，或者晋语区中的并州片）；第三位和第四位代表官话区中的片（比如东北官话中的吉沈片）。需要注意的是，虽然《地图集》对官话、非官话方言区的片进行了细致的划分，我们仅对官话大区的语言片进行了编码，而非官话区的语言片则并没有进行编码。因此，语言编码的后两位数仅在官话区中适用，对于非官话区而言，其方言编码的后两位总是 00。

此外，还有两种特殊情况需要注意。第一种是“一县多码”，也即有的县包含了多于一种方言。在此情况下， Lavely（2012）将其使用的方言在数据中进行列举（至多五种），并标出该地区的“主要方言”。“主要方言”的确定方法如下：当该区县在地图上被某种方言覆盖时，则确定其为主要方言；当该区县在地图上被多于一种方言覆盖时，该县城的方言则被指定为主要方言（在数据集中，主要方言的位置在第一个语言列中；第二列到第四列的方言没有排序）。第二种情况是信息缺失，也即对于不被《地图集》覆盖的区县，对应的方言编码则被留空。

三、 CFPS 方言编码的流程与原则

我们采用了双向独立验证并判定（Two-way Independent Verification with Adjudication）的方式进行编码。第一轮编码由三个编码员分别单独对每一个受访者所填写的方言信息进行编码，若结果一致，则保留；若不一致，则由另一位经验较为丰富的编码员结合 CFPS 数据中的其他信息，重新确定所属编码类别编码。经统计，2012 年成人库中的 QZ104 变量，不同编码员之间的匹配率 83.04%，不匹配的情况在二次编码时结合多变量信息已得到很好解决，可编码的样本达到 99.88%。

编码时，编码员通过被访者填写的文字信息，并结合其所在区县，按照《中国语言地图集》进行编码。整个过程遵循以下基本原则：

a） 受访者的回答为“本地话”：按照其所在区县的方言类型编码

b） 受访者回答出的方言类型与其所在区域的方言不符：以受访者回答为准；

c） 非单一方言及少数民族语言：统一编码为 99（代表无法编码）；

d） 受访者的回答为“家乡话”：参照其出生地及 3 岁时户口所在地信息编码。

四、小结

经过多位编码员的几轮工作，目前已经完成了 2010 年和 2012 年成人库和少儿库中的方言变量的编码工作，具体的工作数量见表 1。2010 年基线调查数据中，成人库中方言变量的数据涉及到 32 个方言片区；少儿库中方言变量的数据涉及到 12 个方言片区。在 2012 年第一轮追踪调查的数据中，成人数据库中的方言类别涉及 69 个方言片区，少儿数据库中的方言种类涉及 66 个方言片区。综合两年的数据，在使用方言的被调查者中，大部分人使用的均为自己家乡的方言。这其中又分为两种情况：第一种，被调查者为当地人，并在家乡当地工作生活，其回答时大部分回答“本地话”或“XX 地区方言”；第二种情况，受访者在家乡以外地区工作生活，此事受访者一般回答“家乡话”。可以看出，受访者所使用的方言，与其成长地域有着明显的关联。

表 1 各数据库方言变量有效编码数据量

变量名	可编码数量	所属数据库
QD2	863	2010 成人
KS3	85	2010 少儿
WK2	116	2010 少儿
QZ104	22046	2012 成人
KZ104	298	2012 少儿
KZ104_B_2	1586	2012 少儿

参考文献

Australian Academy of the Humanities and the Chinese Academy of Social Sciences. 1988. Language Atlas of China. Pacific Linguistics, Series C, No. 102. Hong Kong: Longman Group (Far East) Ltd.

Lavely, William; Berman, Lex, 2012, "Language Atlas of China", http://hdl.handle.net/1902.1/19004 G. W. Skinner Archive [Distributor] V1 [Version]

李荣，中国的语言和方言《中国语言地图集》图[AI]说明稿 [J]. 方言,1989,(3).

中国语言地图集[M]. 朗文出版 (远东) 有限公司, 1987.

附录：

CFPS 方言编码分类

(1)汉藏语系		100000
(1)汉语语族		110000
(1)官话大区		111000
(1)东北官话		111100
	(01)吉沈片	111101
	(02)哈阜片	111102
	(03)黑松片	111103
(2)北京官话		111200
	(01)京师片	111201
	(02)怀承片	111202
	(03)朝峰片	111023
	(04)石克片	111204
(3)冀鲁官话		111300
	(01)保唐片	111301
	(02)石济片	111302
	(03)沧惠片	111303
(4)胶辽官话		111400
	(01)青州片	111401
	(02)登连片	111402
	(03)盖桓片	111403
(5)中原官话		111500
	(01)郑曹片	111501
	(02)蔡鲁片	111502
	(03)洛徐片	111503
	(04)信蚌片	111504
	(05)汾河片	111505
	(06)关中片	111506
	(07)秦陇片	111507
	(08)陇中片	111508
	(09)南疆片	111509
(6)兰银官话		111600
	(01)金城片	111601
	(02)银吴片	111602
	(03)河西片	111603

	(04)塔密片	111604
	(05)北疆片	111605
(7)西南官话		111700
	(01)成渝片	111701
	(02)滇西片	111702
	(03)黔北片	111703
	(04)昆贵片	111704
	(05)灌赤片	111705
	(06)鄂北片	111706
	(07)武天片	111707
	(08)岑江片	111708
	(09)黔南片	111709
	(10)湘南片	111710
	(11)桂柳片	111711
	(12)常鹤片	111712
(8)江淮官话		111800
	(01)洪巢片	111801
	(02)泰如片	111802
	(03)黄孝片	111803
(2)晋语区		112000
	(100)并州片	112100
	(200)吕梁片	112200
	(300)上党片	112300
	(400)五台片	112400
	(500)大包片	112500
	(600)张呼片	112600
	(700)邯新片	112700
	(800)志延片	112800
(3)吴语区		113000
	(100)太湖片	113100
	(200)台州片	113200

	(300)瓯江片	113300
	(400)婺州片	113400
	(500)处衢片	113500
	(600)宣州片	113600
(4)赣语区		114000
	(100)昌靖片	114100
	(200)宜浏片	114200
	(300)吉茶片	114300
	(400)抚广片	114400
	(500)鹰弋片	114500
	(600)大通片	114600
	(700)耒资片	114700
	(800)洞绥片	114800
	(900)怀岳片	114900
(5)湘语区		115000
	(100)长益片	115100
	(200)娄邵片	115200
	(300)吉溆片	115300
(6)闽语区		116000
	(100)闽南区	116100
	(200)莆仙区	116200
	(300)闽东区	116300
	(400)闽北区	116400
	(500)闽中区	116500
	(600)琼文区	116600
	(700)雷州区	116700
	(800)邵将区	116800
(7)粤语区		117000
	(100)广府片	117100
	(200)四邑片	117200
	(300)高阳片	117300

	(400)勾漏片	117400
	(500)吴化片	117500
	(600)邕浔片	117600
	(700)钦廉片	117700
(8)客家话区		118000
	(100)粤台片	118100
	(200)粤中片	118200
	(300)惠州片	118300
	(400)粤北片	118400
	(500)汀州片	118500
	(600)宁龙片	118600
	(700)于桂片	118700
	(800)铜鼓片	118800
(9)徽语区		119000
	(100)旌占片	119100
	(200)绩歙片	119200
	(300)休黟片	119300
	(400)祁德片	119400
	(500)严州片	119500
(0)其他话区		110000
	(100)平话	110100
	(200)儋州话	110200
	(300)乡话	110300
	(400)韶州土话	110400
(2)藏缅语族		120000
(2)南岛语系		200000
(3)阿尔泰语系		300000
(4)南亚语系		400000
(5)印欧语系		500000