Version 1.0, 10/24/00

Coding Scheme for the *Language Atlas of China*

William Lavely
University of Washington
lavely@u.washington.edu

This document describes a coding of Chinese language groups and dialects based on the *Language Atlas of China*, and procedures used in the creation of a county-level data set based on the *Atlas*. As far as we know, this is the first attempt to code Chinese languages and the only systematic digital data set on the distribution of dialects by county in China.

Our motivation for creating the data set was to understand the relationship between cultural variation and demographic behavior in China. Significant links between culture and demography were demonstrated by the Princeton European Fertility Project (Coale and Watkins 1986), which found that the historical fertility transitions occurred within broadly homogeneous linguistic groups. We anticipate that demographic behavior in China has been similarly shaped by culture, as defined by language and dialect. Ideally, individuals, families, or villages would be classified by their language or dialect. However, no data for this exist.

The *Language Atlas of China* is a monumental compilation and generalization of local linguistic studies. The *Atlas* was produced by Chinese and Australian scholars under the auspices of the Australian Academy of the Humanities and the Chinese Academy of Social Sciences, and published in Hong Kong in 1987 by the Longman Group (Far East) Ltd., and the Chinese Academy of Social Sciences. The *Atlas* contains overviews of each language and dialect, as well as maps. Because the information is based on a variety of studies and scholarly sources, the level of detail is inconsistent and the mappings are necessarily highly generalized, as discussed below.

Chinese censuses and surveys collect data on *nationality*, which distinguishes the majority Han Chinese from minority non-Han ethnics. However, sub-ethnic cultural variation among 1,200 million Hans has gone largely unmeasured and unexplored. Dialect is the only plausible indicator that is systematically available. This coding scheme is shaped by our preoccupation with variation among the Han Chinese. It focuses on Han dialects because major non-Han nationalities can be identified directly from census and survey data, and because smaller nationality groups are too sparse for study. Our coding scheme thus identifies non-Han nationality languages only by broad linguistic phyla, and non-Han languages are not coded.

The present coding scheme uses the *Atlas* data to classify *counties* by the language spoken by inhabitants. There are excellent reasons for a county-level coding. First, most of the data in the *Atlas* are based on observations that are identified by county, and usually each dialect is accompanied by a list of counties in which it occurs. In the vast majority of cases, only one dialect is identified per county, and the *Atlas* maps generalize the dialect to the entire county. A county coding is thus consistent with the generalizations made by the *Atlas* compilers. Second, the county is the lowest level administrative unit that is identified by censuses and most surveys. Thus, even if we had finer-grained language data, it would be impossible to link this detail to census and other data. A county-level coding is thus justified by the level of detail provided in the *Atlas* and on practical grounds. Even so, the transposition of the *Atlas* data into a procrustean county framework requires many compromises. Language groupings often cross county boundaries, following terrain contours such as river valleys. Many counties are divided between two or more linguistic groups and thus cannot be coded unambiguously. The best we can do is to indicate those counties where the situation is complex. We have done this by coding every dialect that is mentioned for a county, up to five dialects.

Coding Scheme

The *Atlas* contains an implied coding of the languages of China. It distinguishes six basic levels: phylum, stock, supergroup, group, subgroup, and cluster. The five phyla are: (1) Sino-Tibetan; (2) Austro-Tai; (3) Altaic; (4) Austro-Asiatic, and (5) Indo-European. Within the Sino-Tibetan phylum there are two "stocks," (1) Sinitic and (2) Tibeto-Burman. Our coding scheme follows this design, but because our project is only concerned with Han subethnic variation, only the Sinitic stock under the Sino-Tibetan phylum is coded in detail. Our scheme is designed so that it can be expanded to include non-Chinese languages and all of the *Atlas* detail.

The Sinitic stock is grouped into ten categories, as follows:

- (1) Sino-Tibetan Phylum
  - (1) Sinitic Stock
    - (1) Mandarin supergroup
    - (2) Jin group
    - (3) Wu group
    - (4) Gan group
    - (5) Xiang group
    - (6) Min supergroup
    - (7) Yue group
    - (8) Hakka group
    - (9) Hui group
    - (0) Residual (Pinghua, Danzhou, Xianghua, Shaozhou Tuhua)

This classification follows that used in the *Atlas* very closely (exceptions will be noted below). Note that Mandarin and Min constitute "supergroups" and are listed as if co-equal with groups. Both groups and supergroups define *dialect groups*, that is, groups within which languages are assumed to be mutually intelligible. Speakers of Mandarin dialects, from Manchuria to Yunnan, can generally understand each other (albeit with considerable difficulty). We also make the complementary assumption that *across* groups and supergroups, dialects are mutually unintelligible. These assumptions provide a rationale for classifying the huge (over 600 million speakers) Mandarin group as co-equal with the small (31 million) Xiang group. These assumptions have no bearing on the question of what Chinese people understand as the result of learning other dialects.

For example, the vast majority of Chinese who are non-native speakers of Mandarin understand some Mandarin, if not learned in school, then by exposure to media.

The only modification made to the *Atlas* categories is to relegate four groups (Pinghua, Danzhou, Xianghua, and Shaozhou Tuhua) into a residual category. This has been done due to the very small size of these groups, which are generally only represented in one or two counties.

The next level below the groups and supergroups makes distinctions within these major dialects. Under the "supergroups" are groups, and under the "groups" are subgroups, and, corresponding to the scheme above, "groups" under the supergroups are classed at the same level as "subgroups" of the groups. This is again under the assumption that at this level we are distinguishing among dialects that are at least theoretically mutually intelligible.

The lowest level of code is confined to the Mandarin supergroup. For these groups, we also identify sub-groups, even though it is not clear how different the subgroups are, linguistically or culturally. We know, however, that the Mandarin subgroups are quite large. We have not attempted to capture this level of detail among the non-Mandarin language groups because of their very small size. Some clusters exist in only one county and thus a language effect would be analytically indistinguishable from a county effect.

Coding Procedure

The language code has six digits representing a nested hierarchy:

Phylum (1 digit)
Stock (1)
Supergroup (Mandarin) or Group (1)
Group (Mandarin) (1) or Subgroup (3)
Subgroup (Mandarin only) (2)

Because our interest is only in language variation within Han Chinese, we have coded only the Sinitic Stock of the Sino-Tibetan Phylum. Since the first two digits of the six digit code would always be 11, the first two digits have been dropped.

Data were coded and entered by Ms. Amy Chen, and corrected and verified by Mr.Yong Cai at the University of Washington in summer 2000. The coder was given a list of counties. The *Atlas* provides, for each language subgroup and cluster, a list of counties in which the dialect is spoken. The coder, on the first pass, did not consult maps, but simply coded the counties as listed in the articles provided in the *Atlas*. Some counties are listed under more than one language group. On the second and all subsequent appearances of a county, the coder entered a second variable, third, or higher order variable representing the second, third (etc.) languages, up to a maximum of five. At the end of the first coding pass, the counties with multiple languages were reviewed in light of the *Atlas* maps to determine which dialect is "predominant" in the county. In the majority of cases there is only one language, so the predominant language is unambiguous. For the cases in which more than one language is listed for the county, the language that was mapped in the Atlas was designated the predominant language. In some cases more than one language is mapped for the county. In these cases, the language spoken in the county seat is designated the predominant language. The predominant language was then assigned to the "first" language column. No attempt was made to determine an ordering of the second, third, fourth and fifth languages. Counties that were not mentioned in the Atlas text, and thus were not coded in the first stage of the coding process, were investigated and coded in light of the *Atlas* maps. If the map contained no data, the county received no code. It is apparent that most such counties are in minority areas that speak non-Chinese languages. We found only one county that is unambiguously Han for which the Atlas contained no data, Huangshi Shixiaqu in Hubei (GB 420201).

The *Atlas* contains some inconsistencies in nomenclature. Northern Mandarin, Beifang Mandarin, and Jilu Mandarin refer to the same group.

Data sets

We have produced two data sets, based on two variant administrative codings of Chinese counties in mid-year 1990. The first is the MQ ("merged *qu*") coding, devised by Professor G. William Skinner, UC-Davis, which aggregates small city districts. The MQ coding is designed for mapping and is keyed to a county basemap in ArcView format, also provided (my901.xxx), which was produced by the China In Time And Space project (for documentation, see http://citas.csde.washington.edu/). The second data set is keyed to the coding of counties used by the 1990 census, which is based on (but varies slightly from) the National Standard (*Guobiao*) codes for 1988 (GB 2260-88). The census coding distinguishes city districts. For the purpose of language coding, this detail is spurious, as the *Language Atlas* makes no distinctions among city districts. The present census coding is keyed to the counties represented in a specific 1990 census data set, the 1% clustered sample. The data files are listed in readme.txt, copied at the end of this document.

Use of the Language Data in Analysis

The data files consist of a county identification variable (the GB code), two place name variables, five language variables, and two recodes into broader categories of the predominant language language variable. Only the first language variable is complete because not every county has two Chinese dialects (in fact, in non-Han regions, some have none). As has been noted above, the language code provided (as distinct from the complete code listed below) has only four digits. The first digit is the supergroup or group (e.g., Mandarin, Jin, Wu, etc.); the second digit is the group or subgroup (e.g., Northeastern Mandarin, or Bingzhou in the Jin group); digits 3 and 4 represent the subgroups of Mandarin (e.g., Jishen subgroup of the Northeastern Mandarin group). Note that, the last two digits are germane only to Mandarin dialects. The *Atlas* distinguishes among subgroups of the non-Mandarin dialects, however, this detail is not included in the present coding because of the small size of the groups. For this reason, the last two digits of the non-Mandarin codes are always 00.

Variables in the lang.mq90.dta and lang.cen90.dta files

Variable Name	Variable Label
1. gb	Guobiao code
2. nmcenmq1	Name
3. nmlocal1	Local name
4. _1st	First or predominant language
5. _2nd	Second language
6. _3rd	Third language
7. _4th	Fourth language
8. _5th	Fifth language
9. group10	10 major language groups
10. group18	18 major language groups

The variables "_1st" through "_5th" code language into 43 Mandarin subgroups and 58 non-Mandarin groups, yielding a potential maximum of 101 categories. For many analytic purposes, this detail is unnecessary. Two obvious groupings have been provided here. The variable "group10" codes language into the ten major groups represented by the first digit of the code (Mandarin, Jin, Wu, Gan, etc.). The frequency distribution of counties by "group10" for the merged *qu* county coding is displayed below. Note that 395 counties are "missing" indicating that they are not listed in the *Language Atlas* as speaking a language of Sinitic Stock. These are counties inhabited by non-Han peoples. Note also that 55 percent of counties are classified as Mandarin speaking. The variable "group18" divides the Mandarin Supergroup into 9 major Mandarin groups (Northeastern Mandarin, Beijing Mandarin, etc.) and the 9 major non-Mandarin groups (Jin, Wu, etc.), a total of 18 categories. The recodes provided here are merely two of many reasonable schemes. Another is implied by the first and second digits of the code, which distinguish 9 Mandarin groups and 58 non-Mandarin groups (a total of 67 categories). Other groupings will be guided by empirical findings.

The full coding scheme provided below represents detailed as well as the higher level language categories, thus not every variable contains all of the codes. The variable "group10" for example takes on the values of the higher level codes (1000 for Mandarin, 2000 for Jin Group, etc.), while the variable "_1st" never takes on these values but only assumes the lowest level codes, e.g., 1101 (Jishen), 1102 (Hafu), and 1103 (Heisong) for groups of Northeastern Mandarin. In a very few cases, a county could not be classified to the lowest level, but could be classified at the next higher level. Only in those cases do higher-level codes (ending 00 for Mandarin and ending 000 for non-Mandarin dialects) mix with lower level codes in the same variable.

Frequency Distribution of First Language Categorized by "group10"

Language Group	Counties	Percent
Other Groups	8	.27
Mandarin	1647	55.49
Jin Group	175	5.90
Wu Group	210	7.08
Gan Group	95	3.20
Xiang Group	55	1.85
Min Supergroup	169	5.69
Yue Group	144	4.85
Hakka Group	58	1.95
Hui (Huizhou) Group	12	0.40
. (missing)	395	13.31
Total	2968	100.00

References

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.

Coale, Ansley J, and Susan C. Watkins. 1986. *The decline of fertility in Europe: the revised proceedings of a conference on the Princeton European Fertility Project*. Princeton, New Jersey: Princeton University Press.

GB 2260-88. 1988. 中华人民共和国行政区划代码 [Codes for the administrative divisions of the People's Republic of China]. 国家技术监督局.

Language Codes
Detail for Sinitic stock only*

(1) Sino-Tibetan Phylum	100000
(1) Sinitic Stock	110000
(1) Mandarin supergroup	111000
(1) Northeastern Mandarin	111100
(01) Jishen	111101
(02) Hafu	111102
(03) Heisong	111103
(2) Beijing Mandarin	111200
(01) Jingshi	111201
(02) Huaicheng	111202
(03) Chaofeng	111203
(04) Shike (Beijiang)	111204
(3) Beifang (Jilu) Mandarin	111300
(01) Baotang	111301
(02) Shiji	111302
(03) Canghui	111303
(4) Jiaoliao Mandarin	111400
(01) Qingzhou	111401
(02) Denglian	111402
(03) Gaihuan	111403
(5) Zhongyuan Mandarin	111500
(01) Zhengcao	111501
(02) Cailu	111502
(03) Luoxu	111503
(04) Xinbeng	111504
(05) Fenhe	111505
(06) Guanzhong	111506
(07) Qinlong	111507
(08) Longzhong	111508
(09) Nanjiang	111509
(6) Lanyin Mandarin	111600
(01) Jincheng	111601
(02) Yinwu	111602
(03) Hexi	111603
(04) Tami	111604
(05) Beijiang	111605
(7) Southwestern Mandarin	111700
(01) Chengyu	111701
(02) Dianxi	111702
(03) Qianbei	111703
(04) Kungui	111704
(05) Guanchi	111705

(06) Ebei	111706
(07) Wutian	111707
(08) Cenjiang	111708
(09) Qiannan	111709
(10) Xiangnan	111710
(11) Guiliu	111711
(12) Changhe	111712
(8) Jianghuai Mandarin	111800
(01) Hongchao	111801
(02) Tairu	111802
(03) Huangxiao	111803
(9) Unclassified Mandarin	111900
(2) Jin group	112000
(100) Bingzhou	112100
(200) Luliang	112200
(300) Shangdang	112300
(400) Wutai	112400
(500) Dabao	112500
(600) Zhanghu	112600
(700) Hanxin	112700
(800) Zhiyan	112800
(3) Wu group	113000
(100) Taihu	113100
(200) Taizhou	113200
(300) Oujiang	113300
(400) Wuzhou	113400
(500) Chuqu	113500
(600) Xuanzhou	113600
(4) Gan group	114000
(100) Changjing	114100
(200) Yiliu	114200
(300) Jicha	114300
(400) Fuguang	114400
(500) Yingyi	114500
(600) Datong	114600
(700) Leizi	114700
(800) Dongsui	114800
(900) Huaiyue	114900
(5) Xiang group	115000
(100) Changyi	115100
(200) Loushao	115200
(300) Jixu	115300

(6) Min supergroup	116000
(100) Minnan	116100
(200) Puxian	116200
(300) Mindong	116300
(400) Minbei	116400
(500) Minzhong	116500
(600) Qiongwen	116600
(700) Leizhou	116700
(800) Shaojiang	116800
(7) Yue group	117000
(100) Guangfu	117100
(200) Siyi	117200
(300) Gaoyang	117300
(400) Goulou	117400
(500) Wuhua	117500
(600) Yongxun	117600
(700) Qinlian	117700
(8) Hakka group	118000
(100) Yuetai	118100
(200) Yuezhong	118200
(300) Huizhou	118300
(400) Yuebei	118400
(500) Tingzhou	118500
(600) Ninglong	118600
(700) Yugui	118700
(800) Tonggu	118800
(9) Hui group (Huizhou)	119000
(100) Jingzhan	119100
(200) Jishe	119200
(300) Xiuyi	119300
(400) Qide	119400
(500) Yanzhou	119500
(0) Residual groupings	110000
(100) Pinghua	110100
(200) Danzhou	110200
(300) Xianghua	110300
(400) Shaozhou Tuhua	110400
(2) Tibeto-Burman Stock	120000

(2) Austro-Tai Phylum	200000
(3) Altaic Phylum	300000
(4) Austro-Asiatic Phylum	400000
(5) Indo-European Phylum	500000

*Because detail is provided for Sinitic stock only, the digital data set includes only the last four digits. The first two of the six-digit code (referring to Sinitic stock of the Sino-Tibetan phylum), are always 11, and thus have been omitted.

Readme.txt

Coding Scheme for the Language Atlas of China
Version 1.0, 10/11/00

For access to data files, contact:

William Lavely
University of Washington
lavely@u.washington.edu

Filename	Type	Description
lang.codes.doc	Word2000	Coding scheme and documentation
lang.mq90.dta	Stata6	Merged qu 1990-coded language data
lang.cen90.dta	Stata6	Census 1990-coded language data
my901.shp	ArcView	ArcView shape file
my901.shx	ArcView	ArcView index of feature geometry
my901.dbf	dBase3	County names (ArcView attributes)
lang.mq90.dbf	dBase3	MQ language (ArcView attributes)

Notes on ArcView

The .shp, .shx, and .dbf files together define the geometry and attributes of the counties. These files should be stored in the same project workspace. To join the language data to the ArcView coverage: Open the theme table of my901.shp, and a window named "attributes of my901.shp" will pop up. Click on "gbcenmq" to select a variable for joining with another attribute file, then open language.dbf and select "gb" as the join variable. Then go back to the theme table ("attributes of my901.shp") and click the "join" icon in the menu bar.