Modern Linguistics 现代语言学, 2017, 5(1), 40-57
Published Online February 2017 in Hans. http://www.hanspub.org/journal/ml
https://doi.org/10.12677/ml.2017.51007

Hans汉斯

The North and the South: A Perspective of Their Lexicon and Grammar "Genes"

Mengbing Xiang[1,2]

[1]Department of Chinese Language and Literature & Center for Chinese Linguistics PKU, Key Laboratory of Computational Linguistics, Peking University, Beijing
[2]Department of Chinese Studies, National University of Singapore, Singapore
Email: xiangmb@pku.edu.cn, chsxmen@nus.edu.sg

Received: Feb. 7th, 2017; accepted: Feb. 20th, 2017; published: Feb. 24th, 2017

Abstract

One of the most obvious divisions in Chinese dialects is the confrontation between northern mandarin and southeastern dialects. In this paper, the author selected 16 items from the vocabulary and the grammar volumes of *Linguistic Atlas of Chinese Dialects* and analyzed the feature sequences of the 16 items of 930 Chinese dialects with MEGA (Molecular Evolutionary Genetics Analysis) by simulating DNA sequences. The results showed that lexicon-grammar items alone could also basically reveal the North-South opposition just as phonology items do. Therefore, the introduction of lexicon-grammar items into the Chinese dialect classification is meaningful. Of course, the so-called "feature sequence" in dialectology is not the real DNA sequence and when using MEGA for large sample calculation, it is normal that bootstrap values are low. The important thing is to observe the grouping trends embodied in the phylogenetic trees.

Keywords

Mandarin, Southeastern Dialects, Lexicon-Grammar Feature Sequences, Phylogenetic Analysis

北方官话和东南诸方言——词汇语法"基因"的透视

项梦冰[1,2]

[1]北京大学中文系暨中国语言学研究中心，计算语言学重点实验室，北京
[2]新加坡国立大学中文系，新加坡
Email: xiangmb@pku.edu.cn, chsxmen@nus.edu.sg

收稿日期：2017年2月7日；录用日期：2017年2月20日；发布日期：2017年2月24日

文章引用: 项梦冰. 北方官话和东南诸方言——词汇语法"基因"的透视[J]. 现代语言学, 2017, 5(1): 40-57.
https://doi.org/10.12677/ml.2017.51007

摘　要

汉语方言的一个最明显的分野是北方官话和东南诸方言的对立，即南北对立。本文从《汉语方言地图集》的词汇卷和语法卷选取16个项目，模拟DNA序列用MEGA (Molecular Evolutionary Genetics Analysis 分子进化遗传分析)软件对930个汉语方言点的这16个项目的特征序列进行分析，结果发现只选用词汇、语法项目也能大致看到南北对立，说明在汉语方言的分区工作中引入词汇和语法标准是有意义的。当然方言学里的所谓“特征序列”并非真正的DNA序列，而用MEGA来进行大样本计算时，自展值低也是正常情况，重要的是从中观察进化树所体现的分组趋势。

关键词

官话，东南方言，词汇-语法特征序列，进化分析

1. 引言

汉语方言的一个最明显的分野是北方官话和东南诸方言的对立，即南北对立。本文从《汉语方言地图集》[1]的词汇卷和语法卷选取 16 个词汇、语法项目，模拟 DNA 序列用 MEGA (Molecular Evolutionary Genetics Analysis 分子进化遗传分析)软件进行分析(MEGA 的操作可参看 Hall 2008 [2])，并加以必要的检验，目的是探讨在汉语方言分区研究中引入词汇、语法标准以及借助生物学软件进行辅助分析的可行性。

2. 数据处理

本文选取的 16 个词汇、语法项目在《汉语方言地图集》里的词形分类往往都比较复杂，例如词汇卷 014 图(简称为 LV014)“面儿玉米~，辣椒~”的词形有 4 大类 30 小类，可转写如表 1(略去代表各类词形的符号，“=”表示同音字)。

本文立足于南北对立，将表 1 的 30 种词形概括为两类：“面”及其派生形式为一类，剩下的其他形式为另一类。即 A、C 为一类，B、D 为一类，各有 15 小类。本文只区分大类，不区分小类。本文对选取的 16 个词汇、语法项目所做的词形分类如表 2 所示。“地图集分类”斜线前后的数字分别为《汉语方言地图集》的大类数和小类数。

词形二分法的设计目的是凸显北方的特点。因此凡碰到兼用南北不同词形或南北词形合璧时，都一律归入北方型。例如词汇卷 179“痛~疼”(兼用南北词形)归为“疼”类，而不归为“其他”。词汇卷 101“房子~屋”、“房子~屋子”、“房~屋”、“房~[屋—]”(兼用南北词形)以及“房屋”、“房屋~屋”、“厝~房厝”(南北词形合璧)归为“房”类，而不归为“其他”。

《汉语方言地图集》共设 930 个方言点。根据每个方言点对 16 个词汇语法项目 Y/N 的不同反应，可以得到 930 个 YN 序列。以北京、南京和广州为例(见表 3，第一行为 16 个词汇语法项目的编号)。

即北京、南京、广州的特征序列分别为：YYYYYYYYYYYYYYYY | YYYYYYYYYYNYYYYN | NNNNNYNNNNNNNNNN。其中北京对 16 个项目的反应都是 Y，南京除了对 11、16 两个项目的反应是

Table 1. The classification of word forms for “powder” on Map 014 of LACD vocabulary volume

表 1. LV014 的词形分类

	A	B	C	D
1	面儿	粉	面儿~粉	末
2	面	粉~末	面儿~粉儿	末儿
3	面~面儿	粉~灰	面儿~粉子	末子
4	面子	粉子	面~粉	屑
5	面子~面儿	粉—不包括“粉子”	面面~粉粉	灰
6	面子~面	粉粉		绒
7	面面			糁子
8	面面~面			碴儿
9	面面儿			饽
10	面勇[=]			

Table 2. The word-form classification of 16 lexicon-grammar items
表 2. 16 个词汇、语法项目的词形分类

	地图集编号	地图集分类	本文的二分法
1. 下雨	词汇卷 005	3/13	动词用“下”；动词用其他形式
2. 面儿（玉米~，辣椒~）	词汇卷 014	4/30	“面”类；其他
3. 窝（鸟~）	词汇卷 037	5/23	“窝”类；其他
4. 儿子（叙称）	词汇卷 052	5/61	“儿”类；其他
5. 穿（~鞋）	词汇卷 079	5/9	“穿”类；其他
6. 房子（一座~）	词汇卷 101	5/33	“房”类；其他
7. 锅	词汇卷 109	6/51	“锅”类；其他
8. 站（~起来）	词汇卷 134	4/14	“站”类；其他
9. 走（慢慢儿~）	词汇卷 138	5/15	“走”类；其他
10. 给（他~我一个苹果）	词汇卷 151	6/52	“给”类；其他
11. 疼（摔~了）	词汇卷 179	4/7	“疼”类；其他
12. 他（~姓张）	语法卷 003	6/46	“他”类；其他
13. 不（明天我~去）	语法卷 028	6/31	“不”类；其他
14. 是（他~老师）	语法卷 038	2/5	“是”类；其他
15. 的（我~东西）	语法卷 041	4/7	“的”类；其他
16. 动物性别表示法	语法卷 076	4/17	“公鸡、母鸡”类；其他

Table 3. Examples of feature sequences
表 3. 特征序列举例

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
北京	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
南京	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	N	Y	Y	Y	Y	N
广州	N	N	N	N	N	Y	N	N	N	N	N	N	N	N	N	N

N 外都是 Y，广州只有第 6 个项目的反应是 Y，其他都是 N。每个序列含 16 个 Y/N 信息，930 个序列含 14,880 个 Y/N 信息。这些信息都是笔者目视《汉语方言地图集》手工转写的，做了两遍校对，但个别差错仍或难免，不过当不至影响总体结论。

用人工手段处理 930 个序列不仅耗时费力，还容易出错。因此笔者在逐点查检 16 个项目的 Y/N 信息时即利用 Access2003 直接建立数据库，然后进行归总。930 个序列可以归纳出 240 种不同的序列，各序列所辖的方言点数多寡不同。16 个词汇语法项目的排列顺序完全依据《汉语方言地图集》。如果改变项目的排列顺序，序列的形式就会发生变化，不过不会影响序列种类的数量，因为 930 个方言点对既定项目的 Y/N 反应是相同的。

在 240 种不同的序列里，81 种为官话所独有，不见于东南方言；152 种为东南方言所独有，不见于官话；7 种既见于官话，也见于东南方言。南北共用的 7 种序列本文按官话和东南方言分开，因此共得 247 个序列，见表 4。

表 4 里的名称由方言属性和编号组成，N 代表官话，S 代表东南方言，点数指序列所辖的方言点数，序列由对 16 个词汇语法项目的 Y/N 反应组成，例如 N001-116 表示官话 1 号，辖 116 个方言点，其序列为 YYYYYYYYYYYYYYYY，即对所有的项目都是 Y 反应；S001-084 表示东南方言 1 号，辖 84 个方言点，其序列为 NNNNNNNNNNNNNNN，即对所有的项目都是 N 反应。先排官话独有序列(N001~N081)，然后是南北共用序列中的官话部分(N082~N088)，再后是东南方言独有序列(S001~S152)，最后是南北共用序列中的东南方言部分(S153~S159)。四类序列都一律按所辖方言点数降序排列。南北共用序列的名称都用浅蓝色的字。其中：

N082-002=S159-001 NNNYYNYYYNNYYYYN	N086-001=S153-003 NNNNYNYNNNNNNYNN
N083-002=S156-001 NNNYYNYNYNNYYYYN	N087-001=S157-001 NNNYYNYYYNNYYYYY
N084-001=S154-003 NNNNYNYNYNNYYYYN	N088-001=S158-001 NNYYYNYYYNYYYYYY
N085-001=S155-002 NNNNYNYYYNNYYYYN	

3. 进化分析

为了便于观察，本文先进行小样本量的计算。选取的序列为：N001~N032 (所辖方言点数多于 1 的官话独有序列，辖 306 个方言点)、N082~N088 (见于官话的所有南北共用序列，辖 9 个方言点)、S001~S060 (所辖方言点数多于 1 的东南方言独有序列，辖 462 个方言点)、S153~S159 (见于东南方言的所有南北共用序列，辖 12 个方言点)。总共是 99 种、106 个序列。其中官话 39 个序列，辖 315 个方言点，东南方言 67 个序列，辖 474 个方言点。官话序列的方言点数占官话方言点数的 87% (315/364)，东南方言序列的方言点数占东南方言点数的 84% (474/566)，两者合计占总方言点数的 85% (789/930)。

106 个序列按 DNA 序列的 FASTA 格式进行转写后导入 MEGA (本文用第 6 版)即可进行计算。采用 ME 法(Minimum Evolution 最小进化法)，进行 1000 次构树测试，以 S001-084 为树根(下文的计算都采用同样的方式，不再一一说明)，用曲线形树和环形树两种形式输出进化树，分别如图 1 和图 3 所示。因为输入的序列并非真正的 DNA 序列，而且样本数也比较大，因此本文一律忽略自展值(bootstrap values)，它们实际上都很低。

本文关注的是进化分析所呈现的宏观分组态势，而非序列的具体进化细节。我们用四条彩线在图 1 里分出四个区：(1) A 线区为官话独有序列，高频序列(旁标蓝点者)都集中在这一区；(2) B 线区为官话独有序列和南北共用序列(旁标粉红色的ㄱ)的错杂分布区，以官话独有序列为主；(3) C 线区为东南方言独

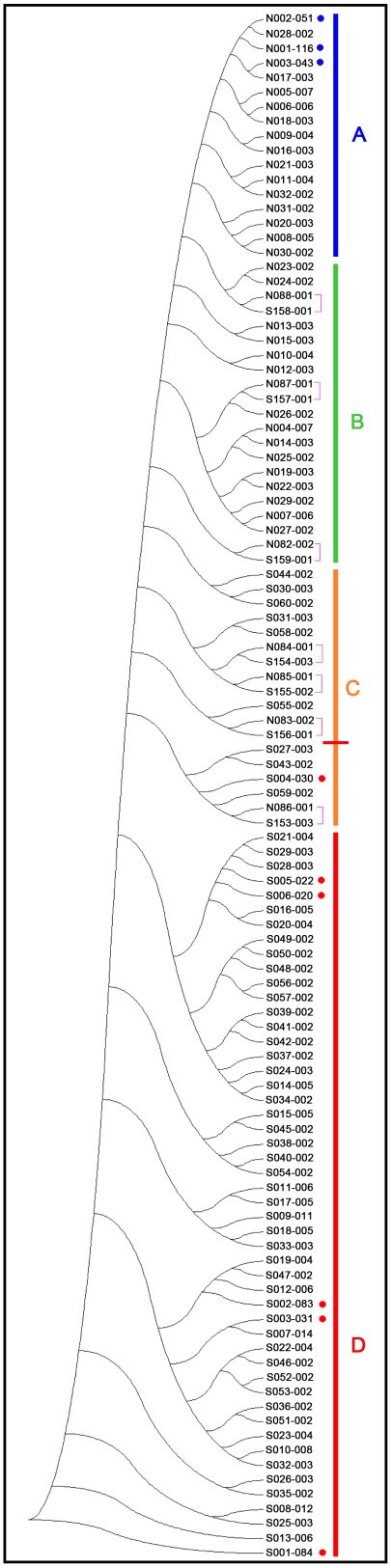

Figure 1. The phylogenetic tree in curve style
图 1. 曲线形进化树

Table 4. All feature sequences
表 4. 特征序列汇总

名称-点数	序列	名称-点数	序列	名称-点数	序列
N001-116	YYYYYYYYYYYYYYYY	N084-001	NNNNYNYNYNNYYYYN	S079-001	NNNNYYNNNNNNNNNN
N002-051	YYYNYYYYYYYYYYYY	N085-001	NNNNYNYYYNNYYYYN	S080-001	NNNNYYYNYNNYNYNN
N003-043	YYYYYNYYYYYYYYYY	N086-001	NNNNYNYNNNNNNYNN	S081-001	NNNYNNNNNNNNNYYY
N004-007	NYYNYYYYYYNYYYYN	N087-001	NNNYYNYYYNNYYYYY	S082-001	NNNYNNNNNYNNNYNY
N005-007	YYYYYYYNYYYYYYYY	N088-001	NNYYYNYYYNYYYYYY	S083-001	NNNYNNNNYNNNNYYN
N006-006	YYYNYYYNYYYYYYYY	S001-084	NNNNNNNNNNNNNNNN	S084-001	NNNYNNNNYNNNYYNN
N007-006	YYYYYYYYYYNYYYYY	S002-083	NNNNNNNNNNNNNYNN	S085-001	NNNYNNNNYNNNYYYN
N008-005	YYYYYYYYYNYYYYYY	S003-031	NNNYNNNNYNNNNYNN	S086-001	NNNYNNYNNNNNYYYN
N009-004	YNYYYNYYYYYYYYYY	S004-030	NNNNNNYNNNNNNYNN	S087-001	NNNYYNNNNNNNNYNY
N010-004	YYNYYYYYYYYYYYYY	S005-022	NNNNYNYNYNNNYYNN	S088-001	NNNYYNNNNNNNYYNN
N011-004	YYYNYYYYYYYYYYYN	S006-020	NNNNNNYNYNNNYYNN	S089-001	NNNYYNNNNNNYNYYN
N012-003	NNNYYYYYNNYYYYYY	S007-014	NNNNNNNNYNNNNYNN	S090-001	NNNYYNNNNYNNNYNY
N013-003	NNYYYNYYYNYYYYYN	S008-012	NNNNNNNNYNNNNNNN	S091-001	NNNYYNNNYNNNNYNY
N014-003	NYYYYYYYYYNYYYYN	S009-011	NNNYYNNNYNNNNYNN	S092-001	NNNYYNNNYYNNNYNN
N015-003	YNYYYNYYYNYYYYYN	S010-008	NYNYNYNNYNNNNYNN	S093-001	NNNYYNYNNNNNNNNN
N016-003	YNYYYYYYYYYYYYYY	S011-006	NNNNYNNNNNNNNNNN	S094-001	NNNYYNYNNNNNYYNY
N017-003	YYYNYNYYYYYYYYYY	S012-006	NNNYNNNNNNNNNYNN	S095-001	NNNYYNYNNNNYNYNN
N018-003	YYYNYYYNYYYNYYYY	S013-006	NYNNNNNNNNNNNNNN	S096-001	NNNYYNYNNNNYNYYN
N019-003	YYYNYYYYYYNYYYYN	S014-005	NNNNNNNNNNNNNYYN	S097-001	NNNYYNYNNYNYNYYN
N020-003	YYYYYNYYYNYYYYYY	S015-005	NNNNNNNNNNNNYYNN	S098-001	NNNYYNYNYNNNNYNY
N021-003	YYYYYNYYYYYYYYYN	S016-005	NNNNNNYNNNNNYYNN	S099-001	NNNYYNYNYNNNYYNN
N022-003	YYYYYYYYYYNYYYYN	S017-005	NNNNYNNNNNNNNYNN	S100-001	NNNYYNYNYNNYYYNN
N023-002	NNYYYNYYYYYYYYYN	S018-005	NNNYYNNNNNNNNYNN	S101-001	NNNYYNYYYNNNYYNN
N024-002	NYYYYNYYYYYYYYYN	S019-004	NNNNNNNNNYNNNYNN	S102-001	NNNYYNYYYNNNYYYY
N025-002	NYYYYYYYYNNYYYYN	S020-004	NNNNYNYNNNNNYYNN	S103-001	NNNYYNYYYYNNYYNN
N026-002	YNYYYNYYYNNYYYYY	S021-004	NNNNYNYYYNNNYYNN	S104-001	NNNYYYNNYNNNNYNN
N027-002	YYYNYYYYYYNYYYYY	S022-004	NNNYNNNNNNNNNYNY	S105-001	NNNYYYNYYNNNNYNN
N028-002	YYYNYYYYYYYYNYYY	S023-004	NNNYNYNNYNNNNYNN	S106-001	NNNYYYYNNNNNNYNN
N029-002	YYYYYNYYYYNYYYYY	S024-003	NNNNNNNNNNNNYYYN	S107-001	NNNYYYYNNNNYYYNY
N030-002	YYYYYYYNYNYYYYYY	S025-003	NNNNNNNNYNNNYNNN	S108-001	NNNYYYYNYNNYYYNN
N031-002	YYYYYYYYYNYYYYYN	S026-003	NNNNNNYNNNNNNNNN	S109-001	NNNYYYYYYNNYNYNN
N032-002	YYYYYYYYYYYYYYYN	S027-003	NNNNNNYNYNNNNYNN	S110-001	NNYNNNNNNNNNNNNN
N033-001	NNNNYNYNNNNNNYYN	S028-003	NNNNYNYNYNNYYYNN	S111-001	NNYNNNNNYYNNNYNN
N034-001	NNNNYNYNYNNNYYYN	S029-003	NNNNYNYYYNNYYYNN	S112-001	NNYNNNNYYNNNNYNN
N035-001	NNNNYNYYYNNYNYYN	S030-003	NNNYYNYYYNNNYYYN	S113-001	NNYNNNYNNNNNNYNN
N036-001	NNNNYYYYNNYYYYYY	S031-003	NNYNYNYNYNNYYYYN	S114-001	NNYNNNYNYNNNNNNN
N037-001	NNNYYYYNNNNYYYYY	S032-003	NYNYNYNNNNNNNYNN	S115-001	NNYNNNYNYYNNYYNN
N038-001	NNNYYYYNYYNYYYYN	S033-003	NYNYYNNNNNNNNYNN	S116-001	NNYNNNYYYNNNYYNN
N039-001	NNNYYYYYNNNYYYYY	S034-002	NNNNNNNNNNNNNNYN	S117-001	NNYNYNNNNNNNNNNN
N040-001	NNYNYNYYNNNYNYYN	S035-002	NNNNNNYNYNNNNNNN	S118-001	NNYNYNNNNNNNNYNN
N041-001	NNYNYNYYYNNYNYYN	S036-002	NNNNNYNNNNNNNYNN	S119-001	NNYNYNNNYNNNNYNN

Continued

N042-001	NNYNYNYYYYNYYYYN	S037-002	NNNNYNNNNNNNNYYN	S120-001	NNYNYNYNNNNYNYNN
N043-001	NNYNYYYYYYNYNYYN	S038-002	NNNNYNNNNNNNYYNN	S121-001	NNYNYNYNNYNNNYYN
N044-001	NNYNYYYYYYNYYYYN	S039-002	NNNNYNNNNNNYNYNN	S122-001	NNYYYNYNNNNYNYNN
N045-001	NNYYYNYYYNNYYYYN	S040-002	NNNNYNNNYNNNYYNN	S123-001	NNYYYNYNYNNNYYNN
N046-001	NNYYYYYYYYNYYYYN	S041-002	NNNNYNYNNNNYNYNN	S124-001	NNYYYNYYYNNNYYYN
N047-001	NYNNYYYYYNNYYYYY	S042-002	NNNNYNYNNNNYYYYN	S125-001	NNYYYNYYYNNYYYYY
N048-001	NYNYYYYYYYNYYYYY	S043-002	NNNNYNYNYNNNNYNN	S126-001	NNYYYNYYYYNYYYYN
N049-001	NYYNYYYNYYNYYYYN	S044-002	NNNNYNYYYNNNYYYY	S127-001	NNYYYNYYYYNYYYYY
N050-001	NYYNYYYYYYNNNYYN	S045-002	NNNYNNNNNNNNYYNN	S128-001	NNYYYYYYYYNYYYYY
N051-001	NYYNYYYYYYNYNYYY	S046-002	NNNYNNNNYNNNNYNY	S129-001	NYNNNNNNNNNNNYNN
N052-001	NYYNYYYYYYNYYYYY	S047-002	NNNYNNNNYYNNNYNN	S130-001	NYNNNYYNYNNYYYYN
N053-001	NYYNYYYYYYYYYYYN	S048-002	NNNYNNYNYNNNNYNN	S131-001	NYNNYNYNNNNYNYNN
N054-001	NYYYYYYYYYNNYYYY	S049-002	NNNYNNYNYNNNYYNN	S132-001	NYNNYNYNNYNYNYYN
N055-001	NYYYYYYYYYNYYYYY	S050-002	NNNYNNYNYNNNYYYN	S133-001	NYNNYNYYNNNYNYNN
N056-001	YNNNYNYYYNNNYYYN	S051-002	NNNYNYNNNNNNNYNN	S134-001	NYNNYNYYYNNYYYNN
N057-001	YNNNYYYNYNNNYYYN	S052-002	NNNYNYNNNNNNNYNY	S135-001	NYNNYNYYYNNYYYYN
N058-001	YNNYNNYNYNYNYYYN	S053-002	NNNYNYNNYNNNNYNY	S136-001	NYNNYYYNNNNNNYNN
N059-001	YNNYYNYNYNYYYYYN	S054-002	NNNYYNNNYNNNYYNN	S137-001	NYNNYYYNNNNYYYNY
N060-001	YNYNYNYYYYNYYYYN	S055-002	NNNYYNYNYNNNYYYN	S138-001	NYNYNNNNNNNNNYNN
N061-001	YNYNYNYYYYYYNYYN	S056-002	NNNYYNYYYNNYNYYN	S139-001	NYNYNNYNYNNNYYYN
N062-001	YNYNYYYNYYYYYYYY	S057-002	NNNYYYYYNYNNYNYN	S140-001	NYNYYYYYYNNYYYNN
N063-001	YNYYYNYYYYNYYYYY	S058-002	NNYNYNYYYNNYYYYN	S141-001	NYYNNNNNNYNNNYNN
N064-001	YNYYYNYYYYYYYYYN	S059-002	NYNNYNYNNNNNNYNN	S142-001	YNNNNNYNYNNNNYYN
N065-001	YNYYYYYYYYNYYYYY	S060-002	YNNYYNYYYNNNYYYN	S143-001	YNNNYNNNNNNNNYNN
N066-001	YNYYYYYYYYYYYYYN	S061-001	NNNNNNNNNNNNNYNY	S144-001	YNNNYNYNNNNNNNNN
N067-001	YYYNYNYNYYYNYYYY	S062-001	NNNNNNNYNNNNYYYN	S145-001	YNNNYNYNYNNNNYYN
N068-001	YYYNYNYNYYYYYYNY	S063-001	NNNNNNNYYYNNNNNN	S146-001	YNNYNNYNNNNYNYNN
N069-001	YYYNYYNNYYYYYYYY	S064-001	NNNNNNYNNNNNYNNN	S147-001	YNNYYNYYYNNNYYYY
N070-001	YYYNYYYNYNYYYYYY	S065-001	NNNNNNYNYNNNYNNN	S148-001	YNYYNNNNNNNYNYYN
N071-001	YYYNYYYYNYNYYYYY	S066-001	NNNNNNYNYYNNNYYN	S149-001	YNYYNNYNNNNYNYYN
N072-001	YYYNYYYYYNNYYYYY	S067-001	NNNNNYNNNNNNNNNN	S150-001	YNYYYNYNYNNYYYYN
N073-001	YYYNYYYYYYNYNYYY	S068-001	NNNNNYYNNNNNNYNN	S151-001	YNYYYNYYYNYYYYYY
N074-001	YYYYYNYNYYYNYYYY	S069-001	NNNNYNNNNNNNNYNY	S152-001	YYNNYNYYYYNNNYYN
N075-001	YYYYYNYNYYYYYYYY	S070-001	NNNNYNNNYNNYYYNN	S153-003	NNNNYNYNNNNNNYNN
N076-001	YYYYYNYYYNYYYYYN	S071-001	NNNNYNNYNNNNNYYN	S154-003	NNNNYNYNYNNYYYYN
N077-001	YYYYYYYNYYYNYYYN	S072-001	NNNNYNNYYNNNNYNN	S155-002	NNNNYNYYYNNYYYYN
N078-001	YYYYYYYYYNYNYYYY	S073-001	NNNNYNNYYNNNYYNN	S156-001	NNNYYNYNYNNYYYYN
N079-001	YYYYYYYYYYNYNYYN	S074-001	NNNNYNYNNNNNYYYN	S157-001	NNNYYNYYYNNYYYYY
N080-001	YYYYYYYYYYNYNYYY	S075-001	NNNNYNYNNNNYYYNN	S158-001	NNYYYNYYYNYYYYYY
N081-001	YYYYYYYYYYYNYYNY	S076-001	NNNNYNYNYNNYNYNN	S159-001	NNNYYNYYYNNYYYYN
N082-002	NNNYYNYYYNNYYYYN	S077-001	NNNNYNYYYNNNNYYNN		
N083-002	NNNYYNYNYNNYYYYN	S078-001	NNNNYNYYYNNNYYYN		

有序列和南北共用序列的错杂分布区，以东南方言独有序列为主；(4) D 线区为东南方言独有序列，高频序列(旁标红点者)主要集中在这一区。如果笼统一点，B 线区和 C 线区也可以概括为一个区，即官话独有序列、东南方言独有序列以及南北共用序列的错杂分布区。图 1 的这种分布态势可以说明官话和东南方言的词汇、语法“基因”大体上是南北有别的。A 线区、B 线区可以归为官话序列(A 线区是典型的，B 线区是非典型的)，C 线区、D 线区可以归为东南方言序列(D 线区是典型的，C 线区是非典型的)。树根(S001-084)是最典型的东南方言序列(即对 16 个词汇语法项目都是 N 反应)，而官话的三个高频序列几乎都集中在树梢，离树根最远。

南北共用序列的存在说明表现为官话序列的方言不一定就是官话，同样，表现为东南方言序列的方言不一定就是东南方言。造成这种情况的主要原因是由方言接触引发的“基因重组”(Gene recombination)。以 N088-001/S158-001 序列 NNYYYNYYYNYYYYYY(B 线区第一个标⊐的序列)为例，它只辖铜陵县(吴语宣州片)、芜湖市(江淮官话洪巢片)两个方言点。宣州片吴语大多处在江淮官话的包围之中，因此铜陵县吴语受江淮官话的冲击导致“基因”变异再明显不过。同样，C 线区的 N085-001/S155-002 序列 NNNNYNYYYNNYYYYN 辖郴州(西南官话)、株洲(湘语长益片)、湘潭县(湘语长益片)三个方言点。郴州所在的西南官话桂柳片湘南小片主要通行于郴州市和桂阳县，北边是赣语耒资片，南边是湘南土话，西南官话桂柳片湘南小片可谓被东南方言南北合围，因此其词汇、语法“基因”向东南方言趋同完全在情理之中。

从图 1 还可以看到，A 线要比 D 线短很多，B 线、C 线则长短大致相当。官话高频序列(116、51、43，占 58%)与低频序列(最大值为 7)落差较大，而且几乎紧挨在一起。东南方言高频序列(84、83、31，30、22、20，占 48%)和低频序列(最大值为 14)的落差较小，分布也不集中。这些情况可以说明：东南方言的“基因变异”范围较大，而且存在明显的非中心化(decentralization)趋势。

从结构上看，图 1 有一个明显不合理的地方。即有一个南北共用序列位于 C 线区的南端(NNNNYNYNNNNNNYNN, N086-001/S153-003)，而在它的北边还有 1 个东南方言的高频序列，即 S004-030(NNNNNNYNNNNNNYNN)。C 线区南端的南北共用序列辖耒阳、洪江、麻阳、凤凰四个方言点。它们的方言归属见表 5(向左的箭头表示同左)。即《汉语方言地图集》[1]承袭《中国语言地图集》[3]的处理，耒阳划归赣语耒资片，洪江、麻阳分归湘语的长益片和娄邵片，凤凰划归西南官话黔北片。到了《中国语言地图集》第二版[4]，除耒阳仍划归赣语耒资片外，其他三个点都有程度不同的改变：洪江划归西南官话湖广片怀玉小片，麻阳、凤凰划归西南官话湖广片湘西小片。洪江、麻阳由划归湘语改成了划归西南官话，凤凰则由西南官话黔北片调整为湖广片湘西小片。

按照图 1 的结构，比较好的办法是不仅洪江、麻阳仍应留在东南方言里，凤凰也应划归东南方言(这里不讨论其具体的归属)。这样一来，图 1 的 D 线就可以一直延伸到 C 线标有红杠的地方(在 S156-001 和 S027-003 之间)，从树根到最远的高频序列这一范围内就不会再出现南北共用序列。事实上，检视凤凰方言的音韵，其白读层所呈现的特性属于东南方言是确凿无疑的(参看项梦冰 2017 [5])。它之所以会被归到西南官话里去，跟《湖南方言调查报告》[6]中的凤凰音系仅仅反映其文读层有很大的关系。

Table 5. The classification of Leiyang, Hongjiang, Mayang, and Fenghuang dialect
表 5. 耒阳、洪江、麻阳、凤凰的方言归属

	中国语言地图集	汉语方言地图集	中国语言地图集第二版
耒阳	赣语耒资片	←	←
洪江	湘语长益片	←	西南官话湖广片怀玉小片
麻阳	湘语娄邵片	←	西南官话湖广片湘西小片
凤凰	西南官话黔北片	←	西南官话湖广片湘西小片

官话、东南方言序列的不同在于：前者 Y 值(Y 的频度值，即每个序列出现 Y 的次数)较高，后者 Y 值较低(参看表 6)。

从方言点数看，A 线区、D 线区是大头，构成了典型的两头大、中间小的哑铃型结构。说明 16 项词汇语法特征对于区分官话和东南方言是基本有效的。或者说，词汇语法的南北“基因”确实是明显有别的。官话、东南方言 Y 值和方言点数的匹配情况(如图 2 所示)也可以直观地说明这一点。

图 3 所呈现的语言景观跟图 1 相同，无需赘述。

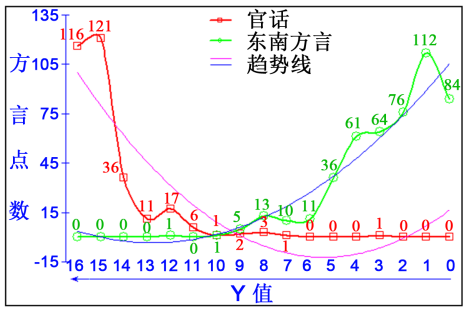

Figure 2. The match between Y frequency values and numbers of dialect locations of northern mandarin and southeastern dialects

图 2. 官话、东南方言 Y 值和方言点数的匹配(106 个序列 789 个方言点)

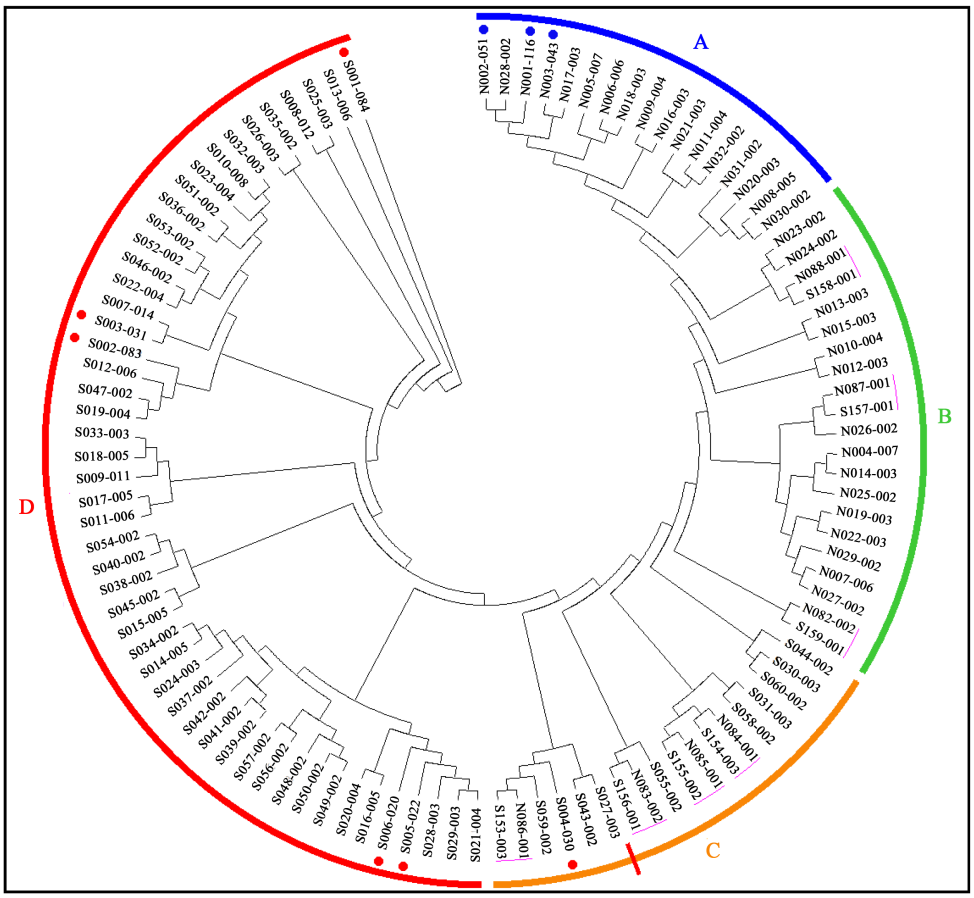

Figure 3. The phylogenetic tree in circle style

图 3. 环形进化树

4. 检验

下面将进行三方面的检验工作。(1) 不同计算方法的对比。(2) 主坐标分析。(3) 247 个序列的完整计算。

4.1. 不同计算方法的对比

MEGA 总共提供了 5 种不同的建树方法。这 5 种方法可分为两组。第一组：1. Maximum Likelihood (最大似然法，简称 ML)。2. Neighbor-Joining (邻接法，简称 NJ)。3. Minimum Evolution (最小进化法，简称 ME)。第二组：4. UPGMA (Unweighted Pair-Group Method with Arithmetic Means，非加权组平均法)。5. Maximum Parsimony (最大简约法，简称 MP)。通常远缘序列选择 ML、NJ、ME，近缘序列选择 UPGMA 或 MP (多用 MP)。

本文对 106 个序列用 5 种方法分别计算了 3 次(其中 MP 在 HP-DX2710SFF-E5200 台式机上计算一次需耗时两个多小时)，然后按计算方法逐一比对 3 棵树，发现即使是同一种计算方法，虽然每次输出的树所反映的大趋势是相同的，但细节并不完全一样(各区段的长短、序列的排列顺序都可能存在不同)。说明序列样本越多，进化路径的可能性越多。或者说，由于自展值非常低，几乎不存在优选项，因此 MEGA 只能从计算结果中随机抓出一棵树来。南北共用序列本文都按方言的性质(官话或东南方言)分开，多数情况下，被人为分开的同一种南北共用序列会聚集在进化树的同一个内部节点(internal node)下，但有时也会被分开。就本文的目的而言，同一种序列会被分开的计算结果自然要加以排除。本文从 ML、NJ、UPGMA、MP 的三次计算中各选一棵树作为对比，见图 4、图 5。

从图 4、图 5 可见，每棵树都有蓝色线段区(A)和红色线段区(D)，官话和东南方言的高频序列分别集中在这两个线段区。ML 树和 MP 树跟图 1 的 ME 树一样，也有绿色线段区(B)和橙色线段区(C)，而 NJ 树和 UPGMA 树则只有紫色线段区(BC)。当然并非 BC 绝对不能分为 B 和 C，而是说其层次不是很分明，勉强分开意义也不大。除了三区(NJ、UPGMA)、四区(ML、MP)的区别外，线段的长短(代表序列的多少)、序列的具体排列顺序每棵树也不一样。如果立足于宏观，可以说图 1、图 4、图 5 所给出的五棵进化树都代表了官话和东南方言大致二分的格局，不同计算方法所得结果的本质并无不同。对比表 6 和表 7 可知，ME 树的层次最为分明(分四区段，各区段的 Y 值表现最符合预期)。

Table 6. The comparison between different sections of the phylogenetic tree
表 6. 进化树不同区段的对比

	序列数量	方言点数	最大 Y 值	最小 Y 值	平均 Y 值
A 线区	17	259	16	13	14.41
B 线区	21	54	15	9	12.14
C 线区	18	65	9	1	5.43
D 线区	50	411	5	0	2.77

Table 7. The comparison between different sections of phylogenetic tree
表 7. 进化树不同区段的对比

	序列数量				方言点数				最大 Y 值				最小 Y 值				平均 Y 值			
	ML	NJ	UP	MP	ML	NJ	UP	MP	ML	NJ	UP	MP	ML	NJ	UP	MP	ML	NJ	UP	MP
A	19	23	25	20	264	274	286	271	16	16	16	16	12	12	12	12	14.16	13.91	14	14
B	13			17	37			41	15			15	11			10	12.85			12.35
BC		38	35			138	84			15	15			2	1			8.47	7.94	
C	45			23	171			110	15			9	1			1	5.96			5.22
D	29	45	46	46	317	377	419	367	5	8	8	9	0	0	0	0	2.86	3.22	3.46	3.85

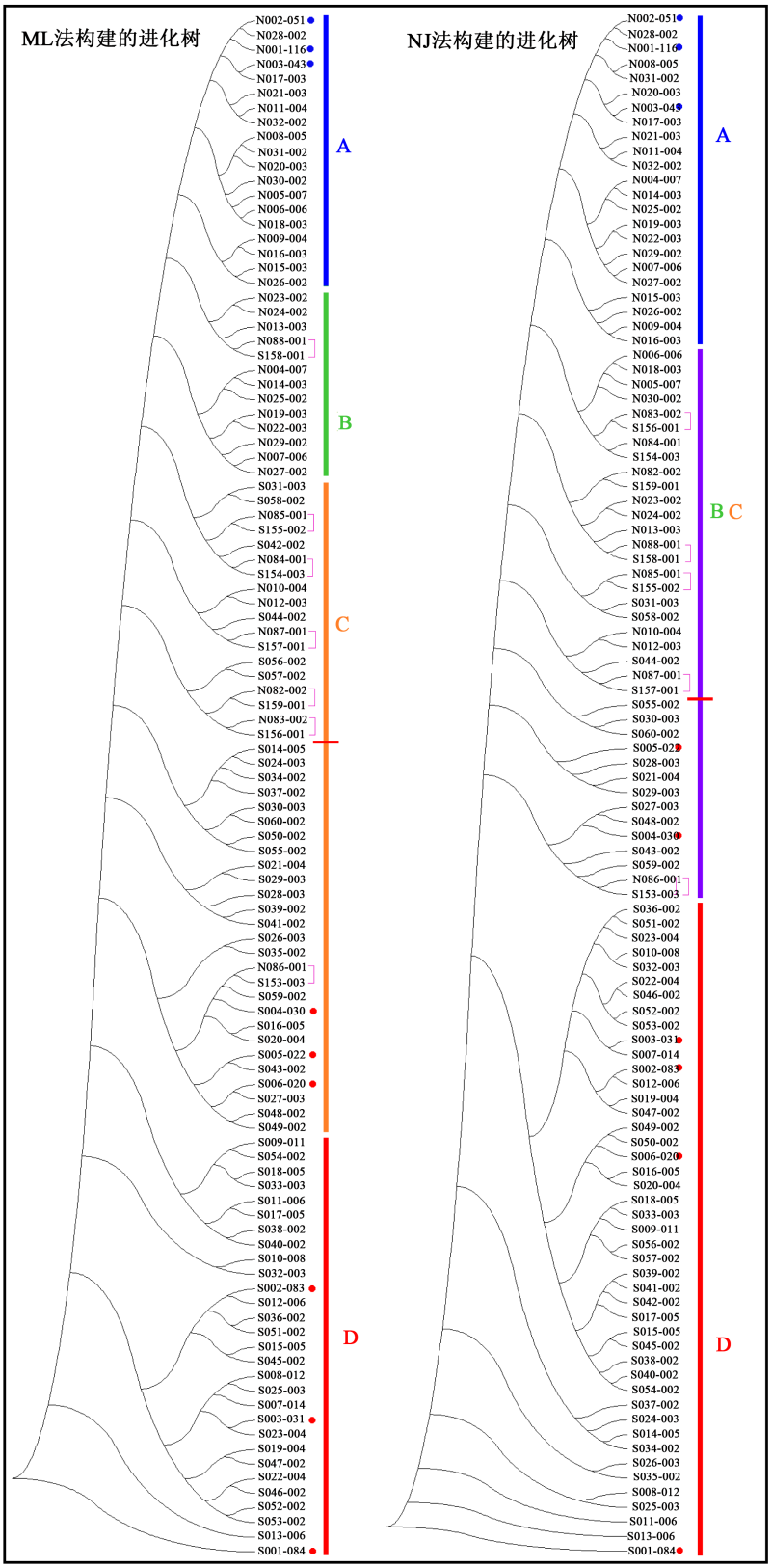

Figure 4. The phylogenetic trees in curve style with ML and NJ methods
图 4. 用 ML 和 NJ 法构建的曲线形进化树

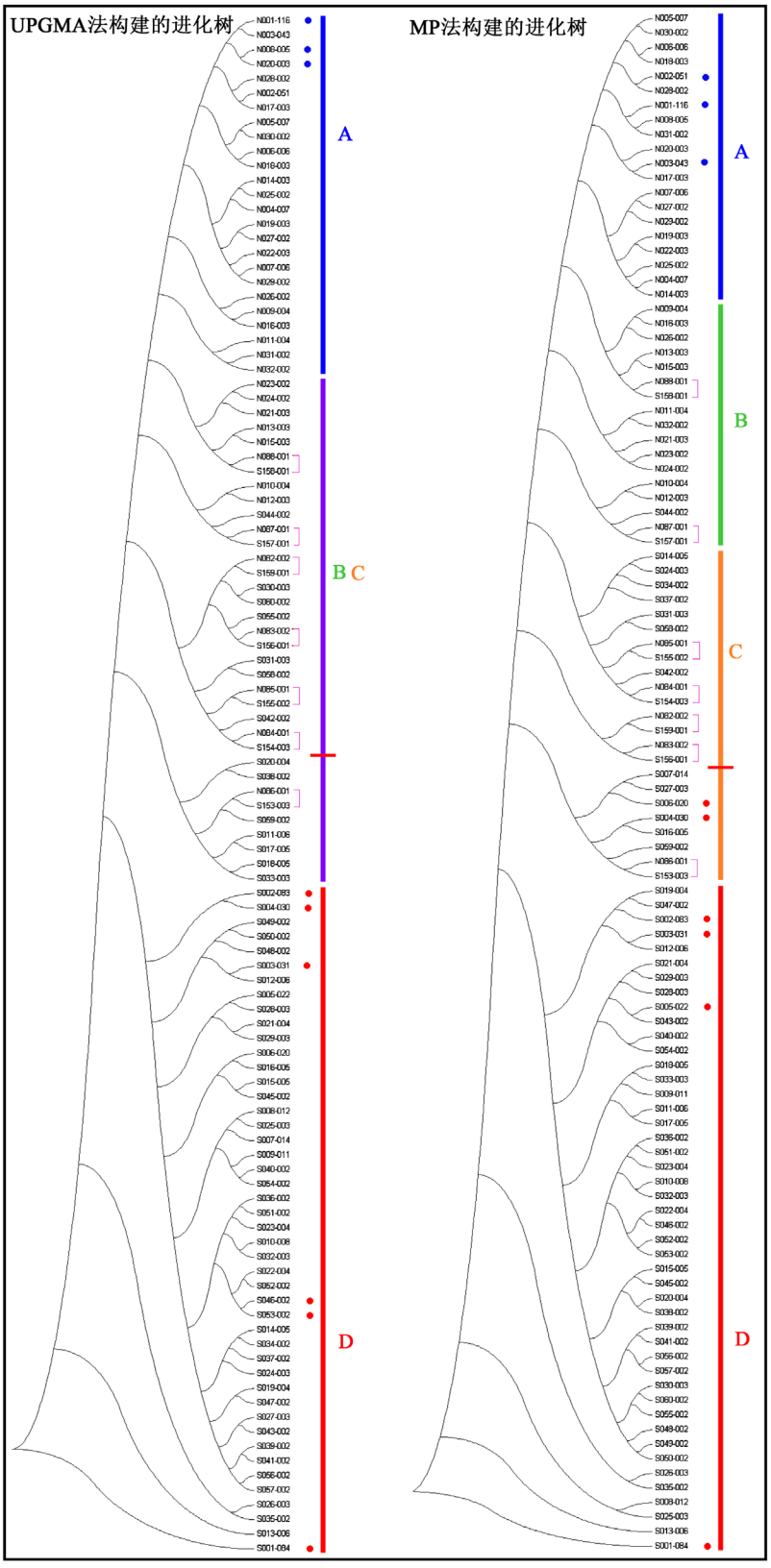

Figure 5. The phylogenetic trees in curve style with UPGMA and MP methods
图 5. 用 UPGMA 和 MP 法构建的曲线形进化树

4.2. 主坐标分析

本文采用 NTSYSpc2.10e 进行主坐标分析，以三维散点图检验 MEGA 的计算结果是否合理。NTSYS 的全称为“数值分类和多元分析系统”(Numerical Taxonomy and Multivariate Analysis System)。NTSYS 在汉语方言学里的运用可参看项梦冰 2015 [7], 2016 [8]。把 106 个序列按 NTSYS 所要求的 0 1 值格式进行转写后进行分步计算，最后可得到如图 4 所示的三维主坐标散点图。官话和东南方言用阿拉伯数字分别标出前三个和前六个高频序列(参看表 4)，南北共用序列则用小写字母标出。由于视角的关系，可能存在一定程度的叠置，因此图中的序列数目也许无法算足 106 个。

从图 6 可见，106 个序列大体在一个半环带上高低错落地分布，一头是纯粹的绿圆(官话序列)，一头是纯粹的红三角(东南方言序列)，高频序列大体都落在半环带的两端，南北共用的序列不仅数量少(约占 7%)，而且主要分布在半环带的中段位置。值得注意的是，尽管 7 个南北共用序列的排列顺序图 6 和图 1 和图 3 不太一样，但两个极端完全相同，即 N086-001、S153-003 (NNNNYNYNNNNNNYNN)在七种南北共用序列的“西端”，N088-001、S158-001 (NNYYYNYYYNYYYYYY)在七种南北共用序列的“东端”。N086-001(凤凰)已经深入东南方言独有序列的“腹地”，前文已指出，它本来就是东南方言。

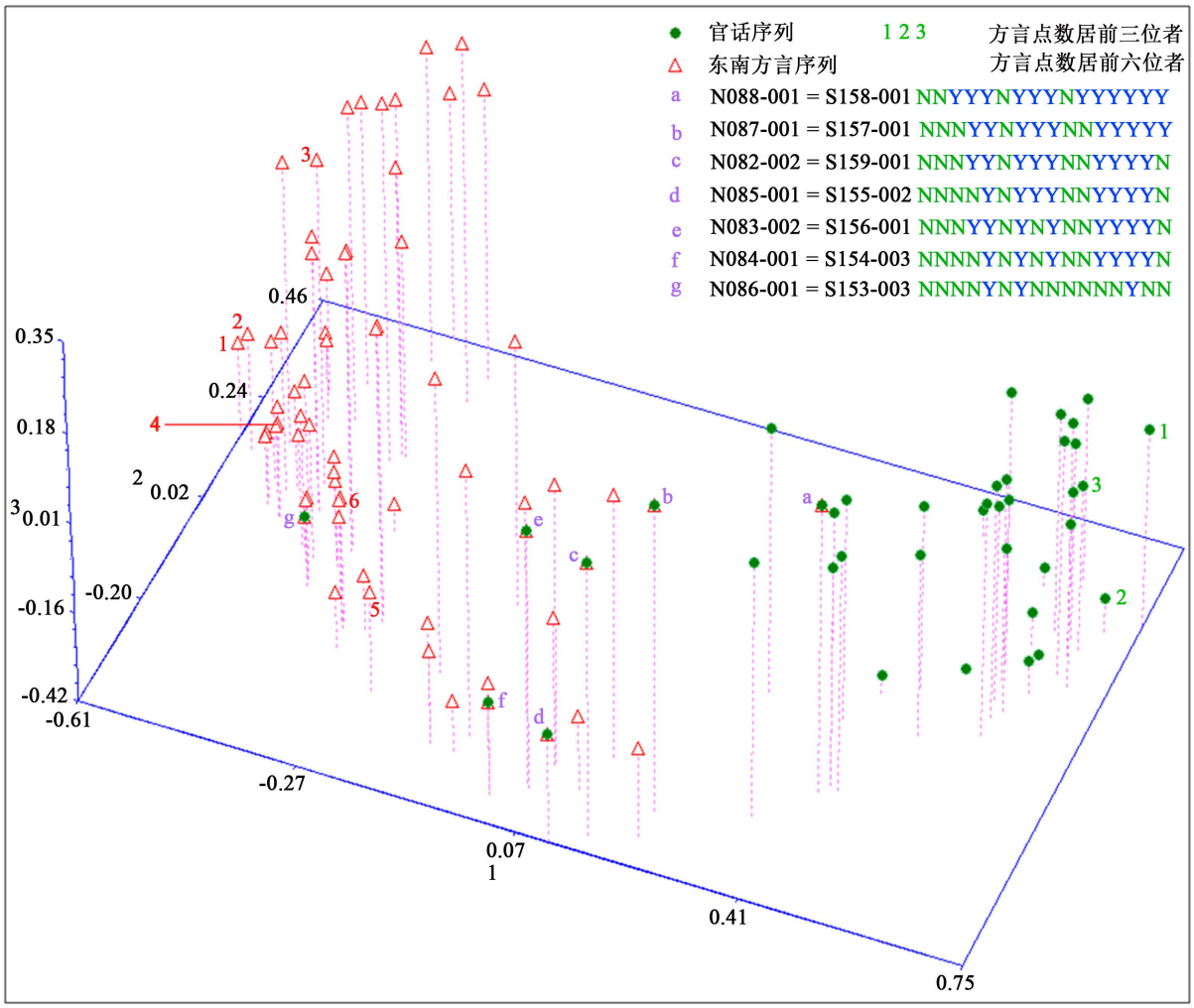

Figure 6. The 3d principal coordinates plot of 106 feature sequences
图 6. 106 个特征序列的三维主坐标图

可见，图 6 和图 1 和图 3 尽管表现形式不同，但所反映的南北序列的分类大势则完全相同。与 MEGA 不同的是，NTSYS 的重复计算结果一定相同。这大概要归因于 MEGA 的进化关系取向和 NTSYS 的相似性取向的不同。在 NTSYS 的计算结果里，南北共用序列必定是叠置的，不会分开。

4.3. 表 4 的 247 个序列的完整计算

前文的计算放弃了官话独有序列和东南方言独有序列中只辖 1 个方言点的序列，分别为 49 种和 92 种，合计 141 种，数量比计算用到的 99 种序列(因为 7 个南北共用序列按方言性质分开实际上是 106 个序列)还要多，因此还需要做完整的计算，看看加入 141 种低频序列后会是一种什么结果。采用 ME 法，计算 4 次，每次约需 3 个小时左右。本文选第 3 次计算结果，以环形图输出。

图 7 和图 3 的结构大致相同，也可分为 4 个明显的区段，南北共用序列分布区的两端依然是 N088、S158 和 N086、S153(分别在 B 区和 C 区)。图 7 的 B 区只出现 1 个南北共用序列，其他 6 个都分布在 C 区，这跟图 3 的 2、5 分配法不同(N086-001、S153-003 实际上不是南北序列)。图 7 中东南方言序列的

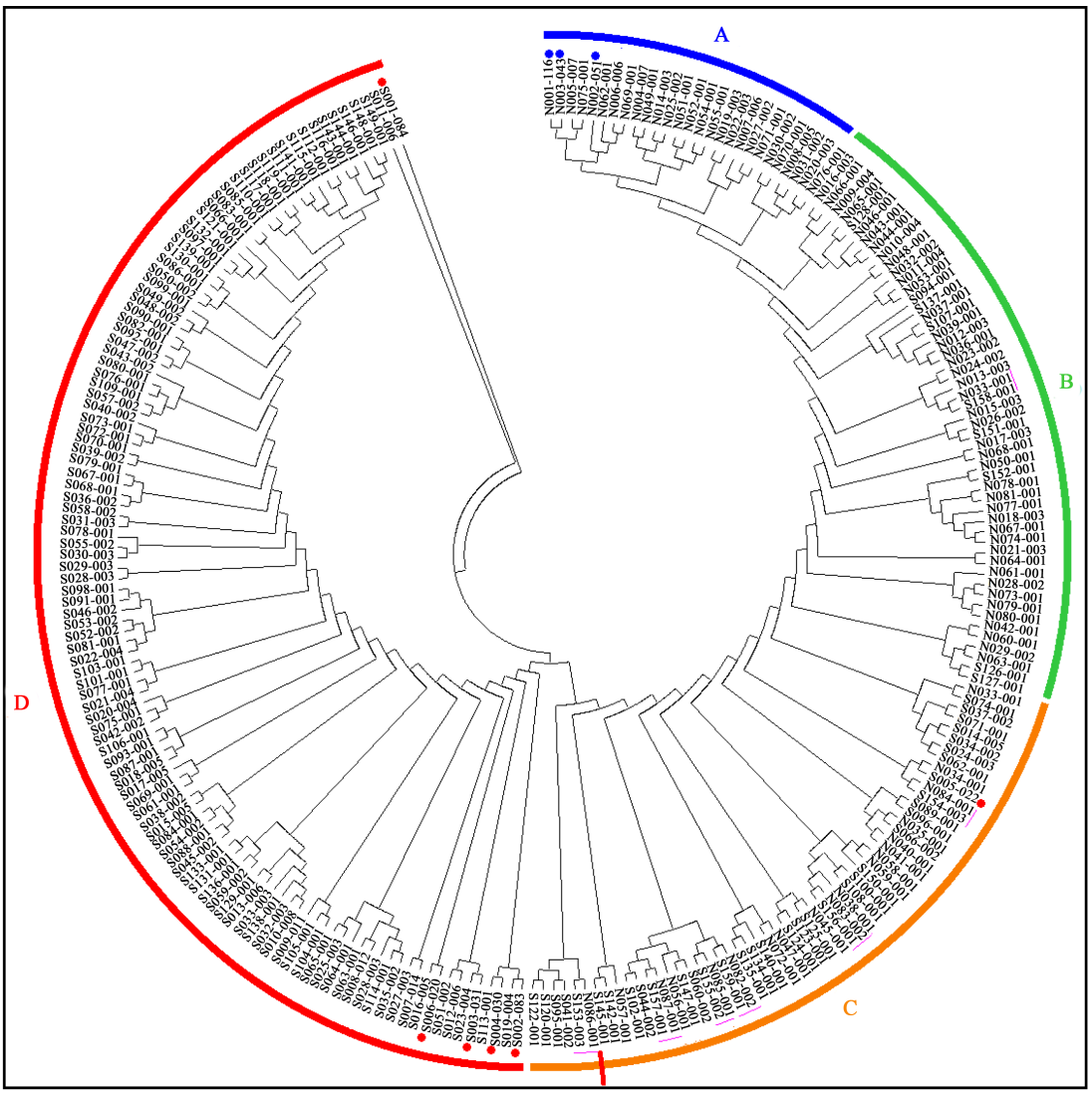

Figure 7. The phylogenetic tree of 247 feature sequences in circle style with ME method
图 7. 用 ME 法构建的 247 个特征序列的环形进化树

非中心化趋势比图 3 更为明显，除 S001 在树根位置外，有 4 个高频序列几乎都集中在 D 区的另一端，而且还有 1 个高频序列(S005-022)深入到 C 区，在 C 区所有南北共用序列之外。可见，加入 141 个低频序列后，进化树的结构合理性会受到一定程度的影响。当然这也取决于我们如何看待高频序列。S005 所辖方言点数仅为 22，跟 S007-014 比只多 8 个方言点，而根据图 7 的表现，我们也可以把 S005 排除在高频序列之外，何况它毕竟还落在 C 区，没有跑到 B 区，并不算太出格。

247 个序列的 Y 值的统计见表 8，Y 值和方言点数的匹配见图 8。表 8、图 8 跟表 6、图 2 所呈现的南北二分景观可以说相当接近。此外，我们还可以从大方言的角度进一步统计出各方言的区段分配和 Y 值情况，如表 9 所示。

Table 8. The comparison between different sections of phylogenetic tree
表 8. 进化树不同区段的对比

	序列数	方言点数	平均每序列方言点数	最大 Y 值	最小 Y 值	平均 Y 值
A	27	272	10.07	16	11	15.13
B	51	80	1.57	15	6	12.60
C	55	86	1.56	13	1	6.43
D	114	492	4.32	9	0	2.50

Table 9. The section distribution and Y frequency value statistics according to the different dialects
表 9. 各方言的区段分配和 Y 值统计

方言		各区段的方言点数				最大 Y 值	最小 Y 值	平均 Y 值	
		A	B	C	D				
官话	兰银	18				16	15	15.78	14.30
	东北	30	2			16	15	15.53	
	北京	9				16	15	15.44	
	胶辽	10	2			16	15	15.42	
	晋陕	39	2			16	13	15.34	
	冀鲁	34	1			16	13	15.20	
	中原	66	16			16	12	14.91	
	西南	50	32	10	1	16	3	12.91	
	江淮	16	16	10		16	6	12.24	
湘语（含乡话）				21	26	9	3	5.83	
赣语			2	26	61	13	2	4.94	
徽语					15	7	2	4.87	
吴语			7	9	105	13	2	4.55	
平话				2	35	7	1	2.70	
土话				1	21	5	0	2.23	
闽语				4	98	4	1	1.38	
粤语				3	57	3	0	0.65	
客家话（含畲话、儋州话）					73	4	0	0.52	

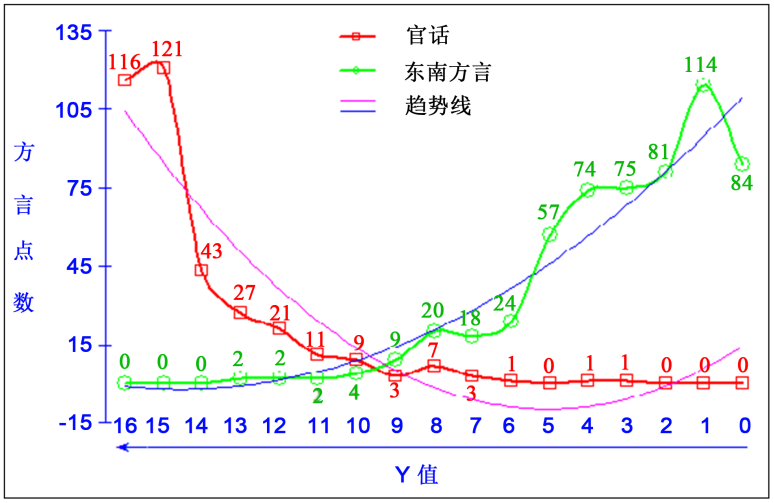

Figure 8. The match between Y frequency values and numbers of dialect locations of northern mandarin and southeastern dialects
图 8. 官话、东南方言 Y 值和方言点数的匹配(247 个序列 930 个方言点)

表 9 以平均 Y 值降序排列，方言分类只是一种便宜的处理，晋陕官话即晋语。兰银、北京官话只分布于 A 区，东北、胶辽、晋陕、冀鲁官话分布于 A 区和 B 区，西南官话四个区都有分布，江淮官话分布于 A、B、C 三区。徽语、客家话只分布于 D 区，湘语、平话、土话、闽语、粤语分布于 C 区和 D 区，赣语、吴语分布于 B、C、D 三区。根据区段分布和平均 Y 值大体可以说：西南官话和江淮官话具有较多的东南方言色彩，而兰银官话至中原官话都是典型的官话；湘、赣、徽、吴具有较多的官话方言色彩，而平话、土话、闽语、粤语、客家话都是典型的东南方言。这种格局显然跟西南官话、江淮官话及湘、赣、徽、吴处在南北交接地带有关。

4.4. 表 4 的 247 个序列的主坐标分析

用 NTSYSpc2.10e 进行主坐标分析得到的三维主坐标图如图 9 所示。

图 9 和图 6 具有完全相同的结构。247 个序列大体在一个半环带上高低错落地分布，一头是纯粹的绿圆(官话序列)，一头是纯粹的红三角(东南方言序列)，排在前三位的序列大体都落在半环带的两端。在半环带的中段，红三角和绿圆或叠合(共有序列)，或穿插，呈现出相当复杂的局面。高频序列、南北共用序列的排列位置跟图 6 完全一样(为了使图面简洁，图 9 东南方言的高频序列只标前三位)。

几乎可以说，图 9 是在不动图 6 的基础上，再把 141 个序列按相对关系一一插入而已。由此可以再一次看到相似性取向和进化关系取向的不同。相似性取向的特点是数据相同、计算方法相同，结果就一定相同。或许有人会因为每一棵进化树可能存在的细节差异而产生不踏实感，可是进化树的好处也是非常明显的：它十分有利于我们做切分，而且可以方便地根据进化树所呈现的面貌做各种数据检验，从而挑选出最好的树。而面对图 9 这样的三维图，不仅切分要如何进行实在有些无计可施，而且标注也势必造成大量的图文叠置而无法观看，不利于各种数据检验。

5. 结论

本文的实验说明，在汉语方言的分区工作中引入词汇和语法标准是有意义的。不过词汇、语法标准的引入并不会根本改变汉语方言分区以音韵标准为主的基本格局。“因为汉语方言的差异以语音为最是一个客观事实，人们历来也是这样看的。这一事实相信将来也不会改变”(王福堂，1999: 46 [9])。在具体

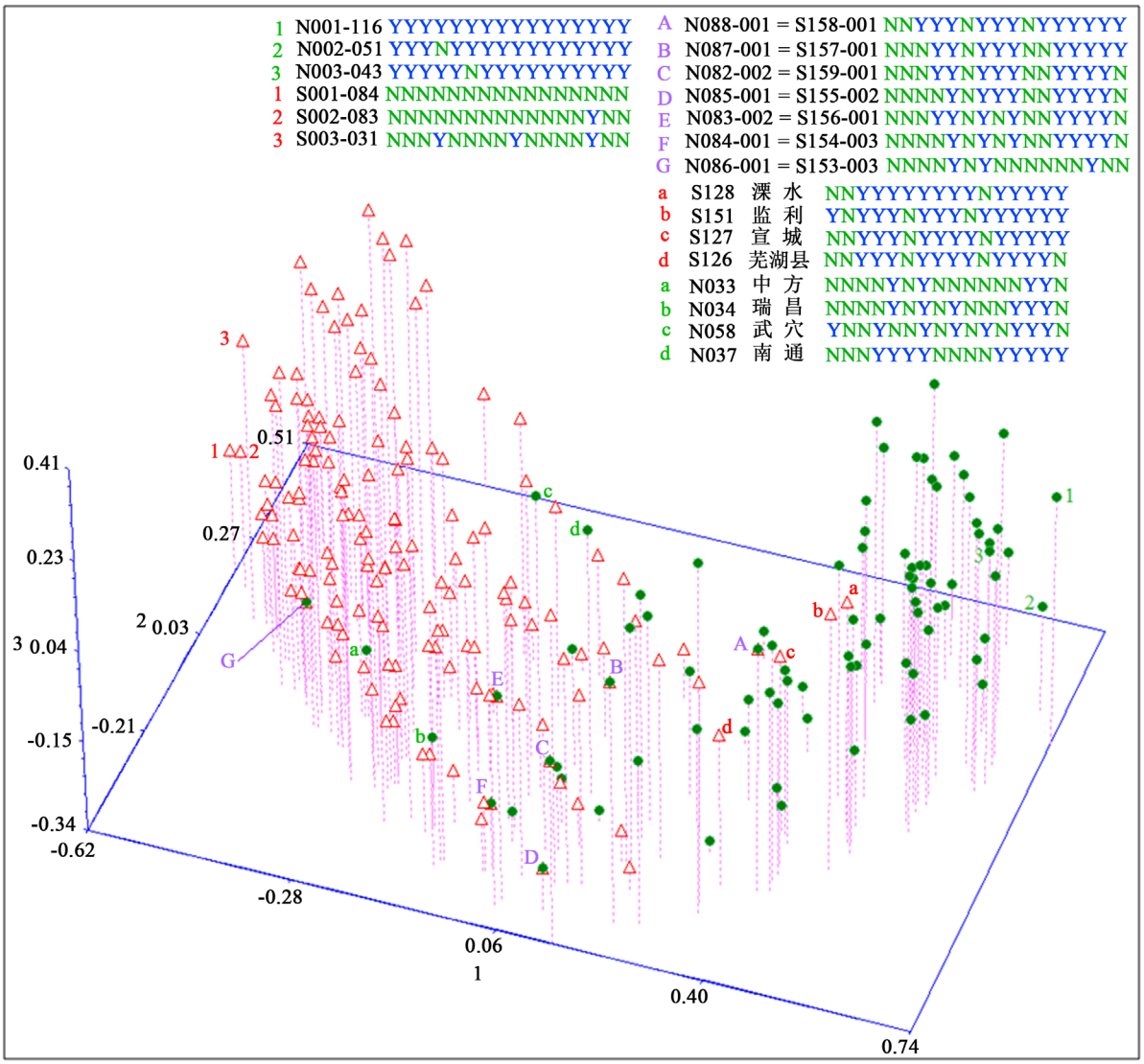

Figure 9. The 3d principal coordinates plot of 247 feature sequences
图 9. 247 个特征序列的三维主坐标图

的操作上，可以把每一个方言对不同分区特征的具体反应比拟为生物学里的一个个 DNA 序列，从而借用生物学的 MEGA 软件来辅助分析。当然方言学里的所谓“特征序列”并非真正的 DNA 序列，而用 MEGA 来进行大样本计算时，自展值低也是正常情况，重要的是从中观察进化树所体现的分组趋势。

致　谢

本文得到国家社科基金重大项目“基于中国语言及方言的语言接触类型和演化建模研究”(项目编号：14ZBD102)的资助，谨致谢忱。

参考文献 (References)

[1] 曹志耘. 汉语方言地图集(分语音、词汇、语法三卷)[M]. 北京: 商务印书馆, 2008.

[2] Hall, B.G. (2008) Phylogenetic Trees Made Easy: A How-To Manual. 3rd Edition, Sinauer Associates, Inc., Sunderland.

[3] 中国社会科学院, 澳大利亚人文科学院. 中国语言地图集[M]. 香港: 朗文出版(远东)有限公司, 1987.

[4] 中国社会科学院语言研究所, 中国社会科学院民族学与人类学研究所, 香港城市大学语言资讯科学研究中心. 中国语言地图集(第二版)[M]. 北京: 商务印书馆, 2012.

[5] 项梦冰. 凤凰方言的归属[J]. 徐州工程学院学报社会科学版, 2017(1).

[6] 杨时逢. 湖南方言调查报告[M]. 台北: 中央研究院历史语言研究所, 1974.

[7] 项梦冰. 聚类分析在汉语方言研究中的运用[J]. 语文研究, 2015(4).

[8] 项梦冰. 古全浊声母的聚类分析和主坐标分析[J]. 云南民族大学学报哲学社会科学版, 2016(3).

[9] 王福堂. 汉语方言语音的演变和层次[M]. 北京: 语文出版社, 1999.