词汇相似度与汉语方言距离计算

Søren Wichmann 冉启斌

ranqibin@126.com

- 北京语言资源高精尖创新中心项目子课题，“语言识别理论及语言数量统计的方法论研究”（KYR17018）之“基于词汇距离计算的语言分类研究”

- 感谢中国社会科学院民族学与人类学研究所江荻研究员对本项目的大力支持！感谢南开大学汉语言文化学院硕士生梁煜珠、孙越，本科生许可等在数据提供与收集方面所做的工作。

- 语言词汇距离的计算：

- ASJP（相似性自动判断程序）数据库： Søren Wichmann等
- http://asjp.clld.org/

- 每种语言收录斯瓦迪士核心词表最稳定的40个词的语音形式

- 7655种语档（doculects）的材料，按ISO639-3有5067种不同的语言

- 示例：dataset.xls；
- listss18.txt

- **距离计算：**
- **Levenshtein编辑距离（LD。1965）：删除，插入，替换**
- **rapt → rap → jap → jaip**
  删除　　替换　　插入
- **3步 → LD = 3**

- LDN距离（LD Normalized。归一化莱文斯坦距离）

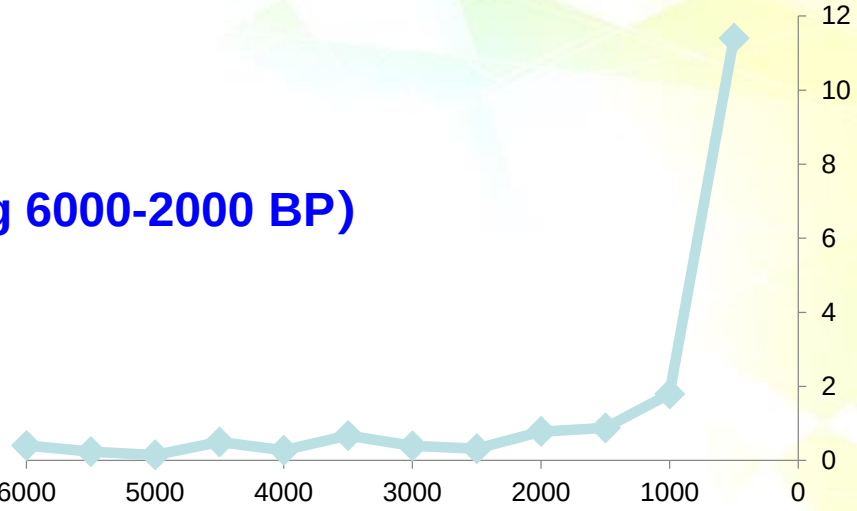
- LDND距离（LDN Divided。
- 归一化莱文斯坦距离商）

$$LDND = \frac{\Sigma_i(d_{ii}) / n}{\sum_{i \neq j}(d_{ij}) / n(n-1)}$$

- 1-LDND：相似度

- **世界语言分类树图**
- **WorldLanguageTree-004（2013）**

- 语言年代学（chronology）
- 语言的迁移速率（~1/3 km/year during 6000-2000 BP）
- 寻找语言分化的发源地点（homeland）
- 史前语言研究（linguistic prehistory）
- 经济学等其他领域
- ……



65种汉语方言数据各方言区分布一览

汉语方言（65）	1. 官话方言（33）	Beijing，Harbin，Shenyang，Ji'nan，阳谷，汶上，聊城，寿光，德州，利津，Rongcheng，Wanrong，Muping，Luoyang，罗山（河南），Xi'an，兰州（甘肃），Yinchuan，Xi'ning，Urumqi，Chengdu，筠连，Guiyang，Kunming，大关（云南），Xuzhou，Wuhan，Nanjing，Guilin，Hefei，Yangzhou，Liuzhou，南通（江苏）
	1. 晋方言（4）	Taiyuan，林州，大同，Xinzhou
	1. 湘方言（3）	Changsha，Loudi，桂阳（湖南）
	1. 赣方言（3）	Nanchang，Pingxiang，Lichuan
	1. 徽方言（1）	Jixi
	1. 吴方言（8）	Shanghai，Chongming，Danyang，Jinhua，Ningbo，Hangzhou，Suzhou，Wenzhou
	1. 闽方言（6）	Fuzhou，Amoy，Chaozhou，Leizhou，Jian'ou，Haikou
	1. 粤方言（5）	Guangzhou，Yangjiang，Nanning，Dongguan，开平
	1. 客家方言（2）	Meixian，Yudu
	1. 平话（0）	--

蒙古
Urumqi
乌鲁木齐
Harbin
哈尔滨
长春
Shenyang
沈阳
Datong
Beijing
呼和浩特
Yinchuan
Taiyuan
Dezhou
Xi'ning
Lanzhou
太原
Wanrong
Wenshang
Xi'an
Luoyang
Xuzhou
西宁
郑州
中华人民共和国
Chengdu
成都
Luosha
Hefei
Shanghai
Wuhan
Jixi
Ningbo
Wenzhou
Daguan
Guiyang
Pingxia
Jian'ou
Fuzhou
Kunming
Guilin
Guiyangdu
Liuzhou
Nanning
Yangjiang
Leizhou
Haikou
昆明
南宁
海口
澳门
香港
拉萨
孟加拉
缅甸
印度
老挝
韩国
日本海
黄海
东海

Shenyang
Datong
Beijing
Yinchuan
Xinzhou
Taiyuan
Dezhou
Lijin
Shouguang
Rongcheng
Xi'ning
Lanzhou
Linzhou
Yanggu
Wenshang
Wanrong
Luoyang
Xi'an
Xuzhou
Luoshan
Hefei
Yangzhou
Danyang
Nantong
Shanghai
Chengdu
Wuhan
Hangzhou
Jixi
Ningbo
Jinhua
Nanchang
Changsha
Wenzhou
Yunlian
Daguan
Loudi
Pingxiang
Lichuan
Jian'ou
Guiyang
Fuzhou
Guiyang
Yudu
Guilin
Kunming
Meixian
Amoy
Liuzhou
Chaozhou
Dongguan
Nanning
Kaiping
Yangjiang
Leizhou
Haikou
- Data © 长地万方 & OpenStreetMap & HERE

- **山西省内方言：**

- **太原，清徐，太谷，祁县，平遥，文水，孝义，平定，和顺，娄烦，离石，临县，汾阳，岚县，隰县，石楼，汾西，五台，代县，五寨，朔州，天镇，山阴，广灵，长治，平顺，长子，沁县，晋城，陵川，阳城，高平，临汾，霍州，闻喜，新绛，运城，吉县，永济**

- **ASJP模式的汉语方言核心词数据库**

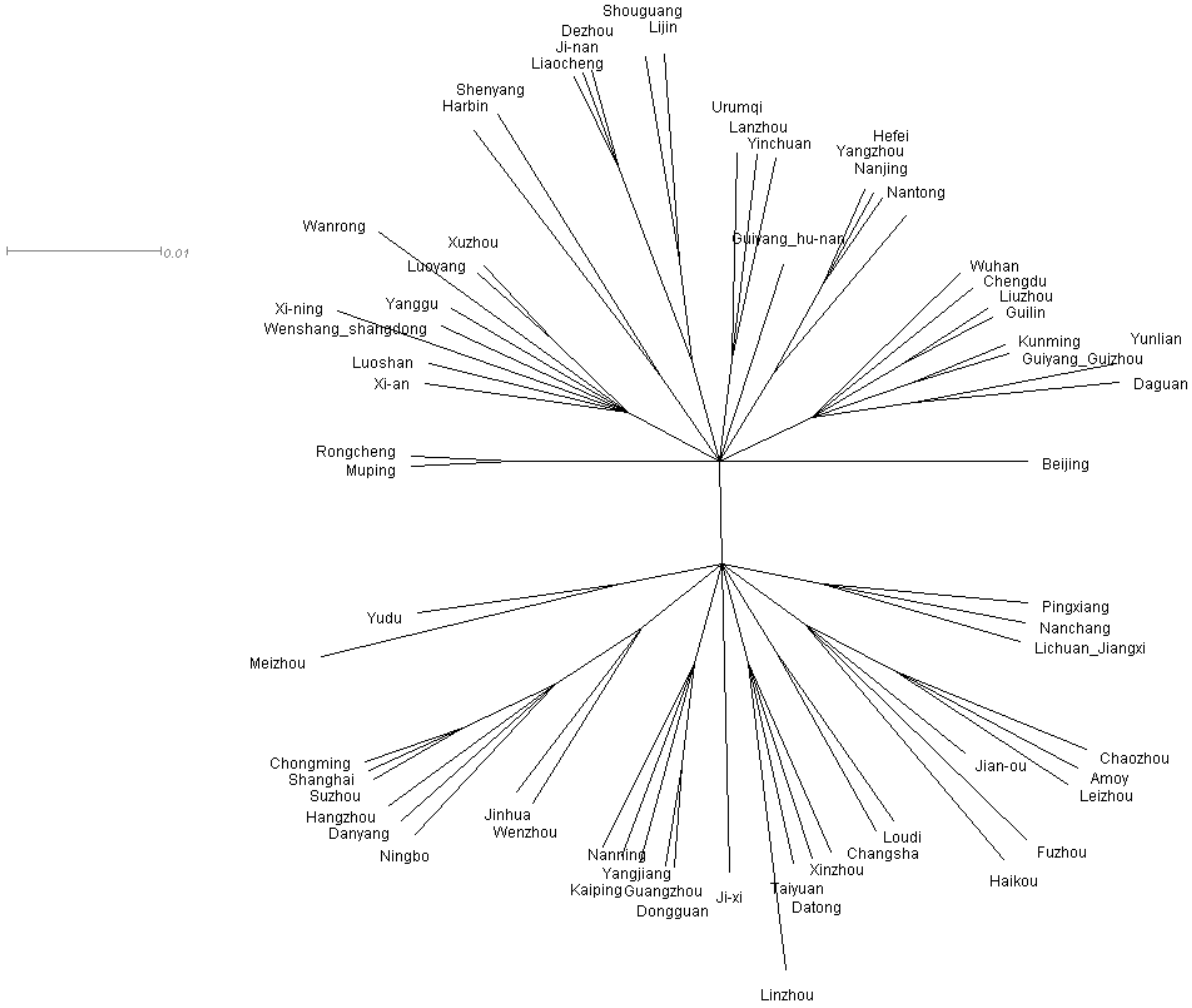
- **65种汉语方言**
- datasetofchinesedialects(20181111).xlsx

- **39种山西方言**
- datasetofshanxi.xlsx

- **”汉语方言核心词语音形式“参考文献.docx**

语言历史发生学树图

- 依据《中国语言地图集》（1987/1990）的分类做出的65种汉语方言历史发生学树图



原ASJP16种汉语方言的词汇距离

2 SYNONYMS, AT LEAST 28 WORDS, DATE > 1700
LOANWORDS EXCLUDED

LDND	AMOY_	CANTO	CHAOY	DONGS	DUNGA	FUZHO	HAINA	HAKKA	KUNMI	MANDA	MANDA	PUTIA	SUZHO	YONGA	XIANG
AMOY_MINNAN_CHINESE{ST.CHI	.00														
CANTONESE{ST.CHINESE|Sino-	88.44	.00													
CHAOYANG_CHINESE{ST.CHINES	58.46	85.59	.00												
DONGSHAN_CHINESE{ST.CHINES	40.40	85.10	35.20	.00											
DUNGAN{ST.CHINESE|Sino-Tib	85.40	87.80	89.30	85.26	.00										
FUZHOU_CHINESE{ST.CHINESE|	74.97	90.24	70.19	69.49	87.37	.00									
HAINAN_MINNAN_CHINESE{ST.C	86.07	90.02	73.84	82.61	93.95	81.71	.00								
HAKKA{ST.CHINESE|Sino-Tibe	83.49	74.78	85.49	80.55	78.54	79.69	84.68	.00							
KUNMING_MANDARIN{ST.CHINES	86.86	89.14	87.85	87.37	64.93	86.39	92.07	78.60	.00						
MANDARIN{ST.CHINESE|Sino-T	89.06	81.94	91.68	89.64	75.48	92.25	94.20	84.15	71.65	.00					
MANDARIN_2{ST.CHINESE|Sino	89.99	86.17	89.17	87.19	73.31	88.59	92.00	77.53	72.32	44.89	.00				
PUTIAN_CHINESE{ST.CHINESE|	67.38	94.39	63.57	60.19	86.84	65.63	83.38	87.00	87.98	93.58	91.94	.00			
SUZHOU_WU{ST.CHINESE|Sino-	88.05	89.89	89.11	85.67	77.71	84.98	95.94	89.59	80.31	86.40	86.05	89.93	.00		
YONGAN_CHINESE{ST.CHINESE|	81.91	89.14	74.59	73.73	83.80	77.74	95.44	84.48	83.99	90.24	86.70	80.74	87.15	.00	
XIANG{ST.CHINESE|Sino-Tibe	85.89	89.59	88.77	84.24	84.03	86.96	96.59	81.38	76.52	86.25	85.71	91.08	80.76	78.96	.00
LDND	AMOY_	CANTO	CHAOY	DONGS	DUNGA	FUZHO	HAINA	HAKKA	KUNMI	MANDA	MANDA	PUTIA	SUZHO	YONGA	XIANG

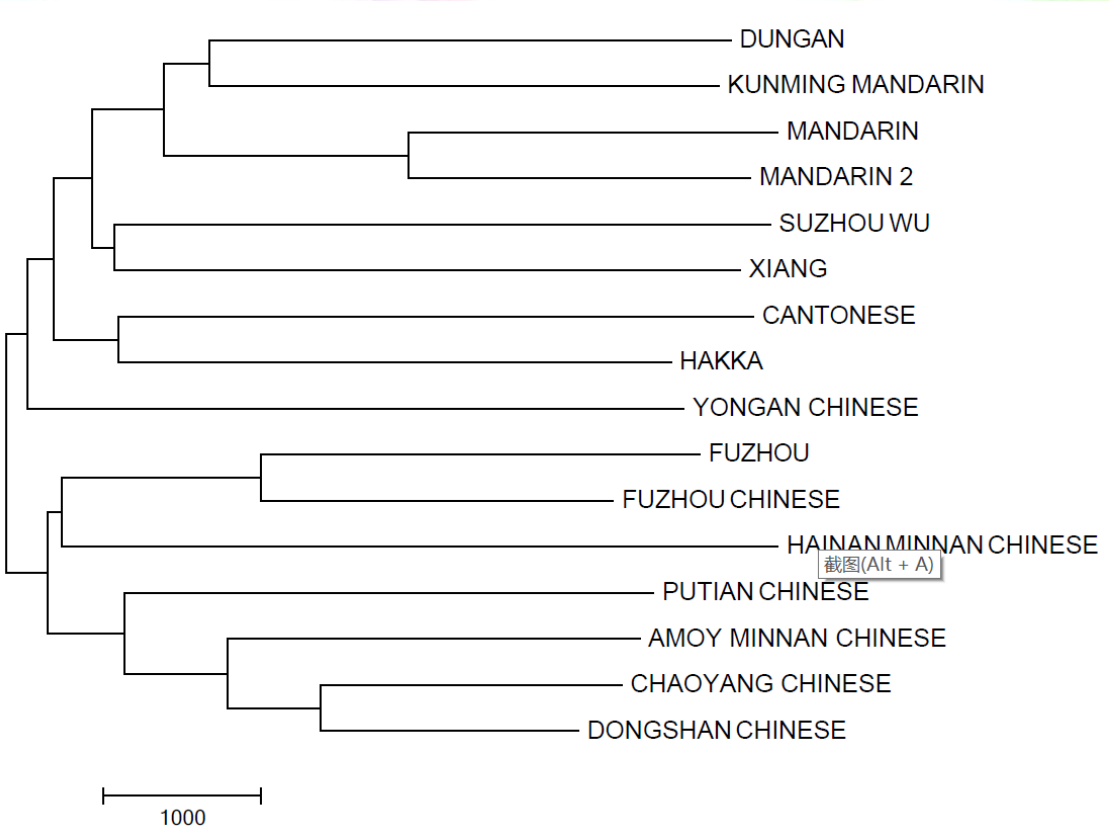
Stop - Program terminated.

原ASJP16种汉语方言的词汇相似度

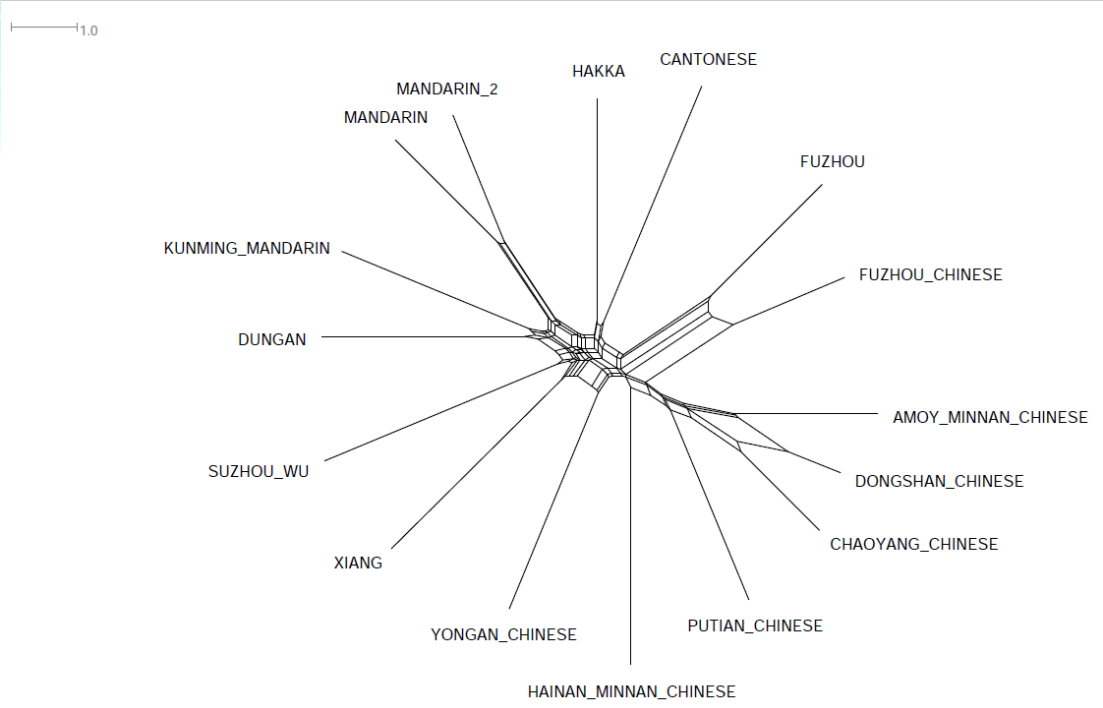
2 SYNONYMS, 28 WORDS, DATE > 1700, WITHIN RANK 3
LOANWORDS EXCLUDED

LANGUAGES		1-LDND%
AMOY_MINNAN_CHINESE	CANTONESE	11.56
AMOY_MINNAN_CHINESE	CHAOYANG_CHINESE	41.54
AMOY_MINNAN_CHINESE	DONGSHAN_CHINESE	59.60
AMOY_MINNAN_CHINESE	DUNGAN	14.60
AMOY_MINNAN_CHINESE	FUZHOU_CHINESE	25.03
AMOY_MINNAN_CHINESE	HAINAN_MINNAN_CHINESE	13.93
AMOY_MINNAN_CHINESE	HAKKA	16.51
AMOY_MINNAN_CHINESE	KUNMING_MANDARIN	13.14
AMOY_MINNAN_CHINESE	MANDARIN	10.94
AMOY_MINNAN_CHINESE	MANDARIN_2	10.01
AMOY_MINNAN_CHINESE	PUTIAN_CHINESE	32.62
AMOY_MINNAN_CHINESE	SUZHOU_WU	11.95
AMOY_MINNAN_CHINESE	YONGAN_CHINESE	18.09
AMOY_MINNAN_CHINESE	XIANG	14.11
CANTONESE	CHAOYANG_CHINESE	14.41
CANTONESE	DONGSHAN_CHINESE	14.90
CANTONESE	DUNGAN	12.20
CANTONESE	FUZHOU_CHINESE	9.76
CANTONESE	HAINAN_MINNAN_CHINESE	9.98
CANTONESE	HAKKA	25.22
CANTONESE	KUNMING_MANDARIN	10.86
CANTONESE	MANDARIN	18.06
CANTONESE	MANDARIN_2	13.83
CANTONESE	PUTIAN_CHINESE	5.61
CANTONESE	SUZHOU_WU	10.11
CANTONESE	YONGAN_CHINESE	10.86
CANTONESE	XIANG	10.41
CHAOYANG_CHINESE	DONGSHAN_CHINESE	64.80
CHAOYANG_CHINESE	DUNGAN	10.70
CHAOYANG_CHINESE	FUZHOU_CHINESE	29.81
CHAOYANG_CHINESE	HAINAN_MINNAN_CHINESE	26.16
CHAOYANG_CHINESE	HAKKA	14.51
CHAOYANG_CHINESE	KUNMING_MANDARIN	12.15

基于词汇距离的原ASJP16种汉语方言分类



基于词汇距离的原ASJP16种汉语方言分化树图



- 目标：以汉语方言为主，达到200个地点的方言

- 考察的内容：
- 语言年代学（chronology）

- 语言/方言的迁移速率

- 语言/方言分化的发源地点（homeland）

- 史前语言研究（linguistic prehistory）
- ……

謝謝