Modelling Chinese dialect evolution

Johann-Mattis List, Shijulal Nelson-Sathi, Dagan Tal

Modelling Chinese Dialect Evolution

Johann-Mattis List[*], Shijulal Nelson-Sathi[+], and Tal Dagan[+]

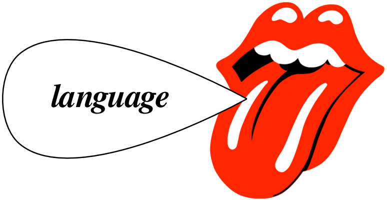
[*]Institute for Romance Languages and Literature
[+]Institute for Genomic Microbiology
Heinrich Heine University Düsseldorf

2012/08/31

Structure of the Talk

1. Languages
   - Languages
   - Diasystems
   - Change
2. Modelling Language History
   - Trees
   - Waves
   - Networks
3. Modelling Chinese Dialect History
   - Data
   - Analysis
   - Results

Languages



Languages and Dialects

Norwegian, Danish, and Swedish are different languages.

Beijing-Chinese, Shanghai-Chinese, and Hakka-Chinese are dialects of the same Chinese language.

Languages and Dialects

Beijing Chinese	1	iou^{21}	i^{55}	xuei^{35}	pei^{21}fəŋ^{55}	kən^{55}	tʰai^{51}iaŋ^{11}	ʈ͡ʂəŋ^{55} ʦai^{53}	naɚ^{51}	ʈ͡ʂəŋ^{55}luən^{51}
Hakka Chinese	1	iu^{33}	it^{55}	pai^{33}a^{11}	pet^{33}fuŋ^{33}	tʰuŋ^{11}	ɲit^{11}tʰeu^{11}	hɔk^{33}	e^{53}	au^{55}
Shanghai Chinese	1	ɦi^{22}		tʰɑ̃^{55} ʦɿ^{21}	poʔ^{3}foŋ^{44}	taʔ^{5}	tʰa^{33}ɦiã^{44}	ʦəŋ^{33} hɔ^{44}		ləʔ^{1}lə^{23}ʦa^{53}

Beijing Chinese	2	ʂei^{35}		də^{55}		pən^{35} liŋ^{21}	ta^{51}
Hakka Chinese	2	man^{33}	ɲin^{11}		kʷɔ^{55}	vɔi^{53}	
Shanghai Chinese	2	sa^{33}	ɲiŋ^{55}	ɦəʔ^{21}		pəŋ^{33} zɿ^{44}	du^{13}

Norwegian	1	nuːɾɑʋinˑn̩	ɔ	suːln̩					kɾɑŋlət	ɔm
Swedish	1	nuːɖanvɪndən	ɔ	suːlən		ty̥ɪstadə	ən gɔŋ			ɔm
Danish	1	noʌ̯ʌnvenˀn̩	ʌ	soːl̩ˀn	kʰʌm		eŋɡ̊ɑŋ	i sd̥ʁiðˀ		ʌmˀ

Norwegian	2	ʋem	ɑ	dem	sm̩	ʋɑː	ɖɳ̩	stæɾ̥kəstə
Swedish	2	vɛm	ɑv	dɔm	sɔm	vɑ		staɹkast
Danish	2	vɛmˀ	a	b̥m̩	d̥	vɑ	d̥n̩	sd̥æʌ̯ɡ̊əsd̥ə

Languages and Dialects

From the perspective of the lexicon and the sound system, the Chinese **dialects** are at least equally if not more different than the Scandinavian **languages**.

Language as a Diasystem

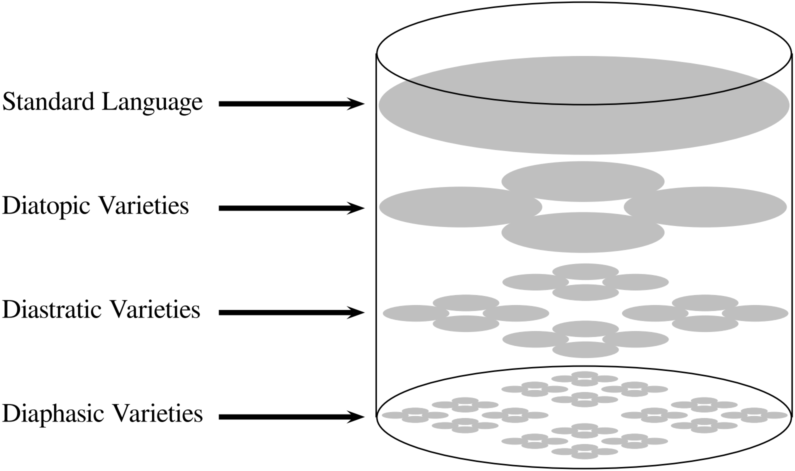
Languages are complex aggregates of different linguistic systems that 'coexist and influence each other' (Coseriu 1973: 40, my translation).

Language as a Diasystem

Languages are complex aggregates of different linguistic systems that 'coexist and influence each other' (Coseriu 1973: 40, my translation).

A linguistic diasystem requires a "roof language" (Goossens 1973:11), i.e. a linguistic variety that serves as a standard for interdialectal communication.

Language as a Diasystem



Standard Language

Diatopic Varieties

Diastratic Varieties

Diaphasic Varieties

Change

Change

expected Mandarin [ma$_{55}$po$_{21}$lou]

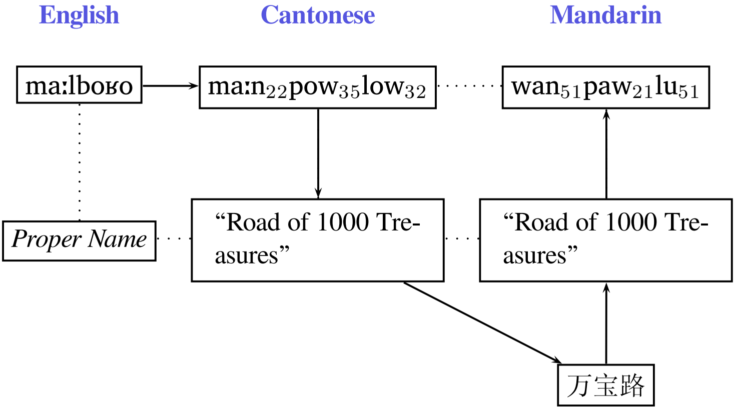
Change

expected	Mandarin	$[\text{ma}_{55}\text{po}_{21}\text{lou}]$
attested	Mandarin	$[\text{wan}_{51}\text{paw}_{21}\text{lu}_{51}]$

Change

expected	Mandarin	[ma$_{55}$po$_{21}$lou]
attested	Mandarin	[wan$_{51}$paw$_{21}$lu$_{51}$]
explanation	Cantonese	[maːn$_{22}$pow$_{35}$low$_{32}$]

Change



English **Cantonese** **Mandarin**

maːlboʁo

maːn$_{22}$pow$_{35}$low$_{32}$

wan$_{51}$paw$_{21}$lu$_{51}$

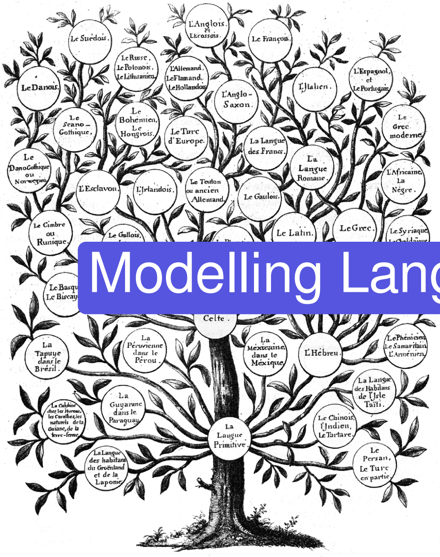
Proper Name

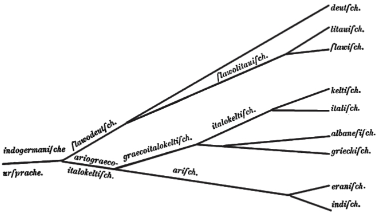
“Road of 1000 Treasures”

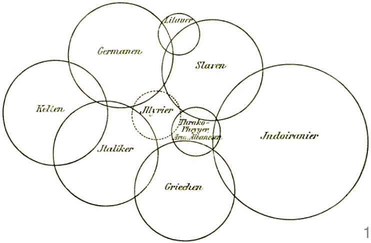
“Road of 1000 Treasures”

万宝路

Modelling Language History


Le Suédois.
L'Anglois. et L'Ecossois.
Le François.
Le Danois.
Le Russe, Le Polonois, Le Lithuanien.
L'Allemand, Le Flamand, Le Hollandois
L'Anglo-Saxon.
L'Italien.
L'Espagnol et Le Portugais.
Le Scano-Gothique.
Le Bohémien, Le Hongrois.
Le Turc d'Europe.
La Langue des Francs.
Le Grec moderne.
Le DanoGothique ou Norwegien
L'Esclavon.
L'Irlandois.
Le Teuton ou ancien Allemand.
Le Gaulois.
La Langue Romane.
L'Africaine La Négre.
Le Cimbre ou Runique.
Le Gallois.
Le Latin.
Le Grec.
Le Syriaque
Le Basque Le Biscayen
Celte.
La Tapuye dans le Brésil.
La Péruvienne dans le Pérou.
La Mexicaine dans le Méxique.
L'Hébreu.
Le Phénicien Le Samaritain L'Arménien.
La Langue des Habitans de l'Isle Taïti.
La Caraïbe chez les Hurons, les Caraïbes, les naturels de la Guiane, de la terre-ferme.
La Guyarane dans le Paraguay.
Le Chinois, l'Indien, Le Tartare.
La Langue Primitive.
La Langue des habitans du Groënland et de la Laponie.
Le Persan, Le Turc en partie.





Dendrophilia

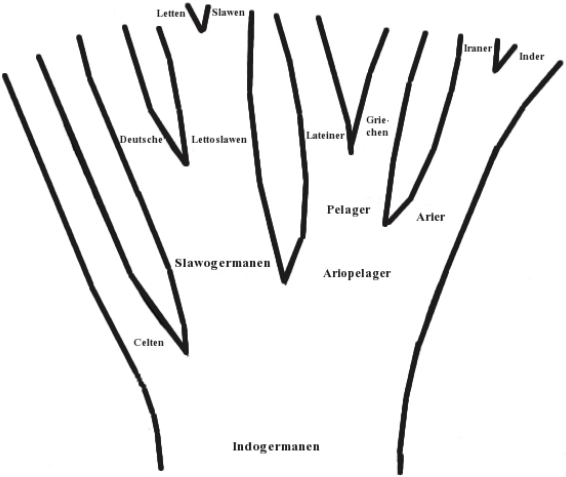
August Schleicher
(1821-1868)

Dendrophilia

These assumptions that logically follow from the results of our research can be best illustrated with help of a branching tree. (Schleicher 1853: 787, my translation)

August Schleicher (1821-1868)

Dendrophilia



Schleicher (1853)

Dendrophobia

Johannes Schmidt
(1843-1901)

Dendrophobia

No matter how we look at it, as long as we stick to the assumption that today's languages originated from their common proto-language via multiple furcation, we will never be able to explain all facts in a scientifically adequate way. (Schmidt 1872: 17, my translation)

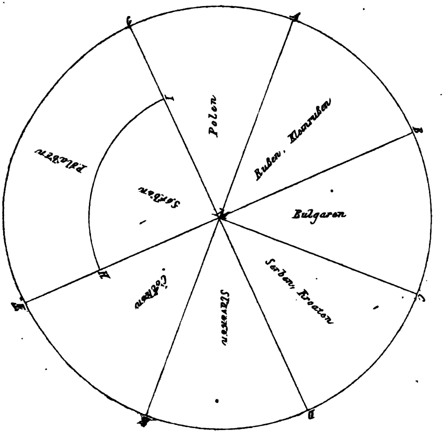
Johannes Schmidt (1843-1901)

Dendrophobia

I want to replace [the tree] *by the image of a wave that spreads out from the center in concentric circles becoming weaker and weaker the farther they get away from the center.* (Schmidt 1872: 27, my translation)

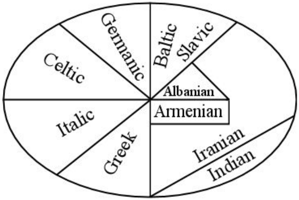
Johannes Schmidt (1843-1901)

Dendrophobia

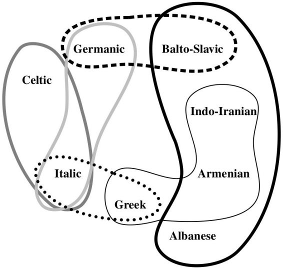


Schmidt (1875)

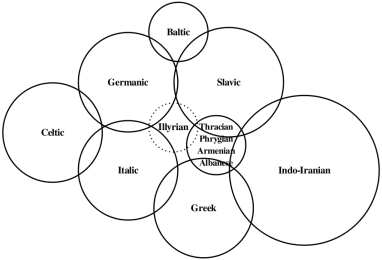
Dendrophobia



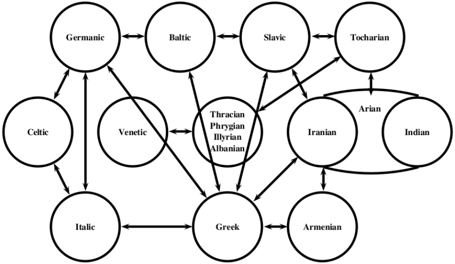
Meillet (1908)



Bloomfield (1933)



Hirt (1905)



Bonfante (1931)

Phylogenetic Networks

Trees are bad because

Phylogenetic Networks

Trees are bad because

- they are difficult to reconstruct

Phylogenetic Networks

Trees are bad because

- they are difficult to reconstruct
- languages do not separate in split processes

Phylogenetic Networks

Trees are bad because

- they are difficult to reconstruct
- languages do not separate in split processes
- they are boring, since they only capture certain aspects of language history, namely the vertical relations

Phylogenetic Networks

Trees are bad because

- they are difficult to reconstruct
- languages do not separate in split processes
- they are boring, since they only capture certain aspects of language history, namely the vertical relations

Waves are bad because

- nobody knows how to reconstruct them

Phylogenetic Networks

Trees are bad because

- they are difficult to reconstruct
- languages do not separate in split processes
- they are boring, since they only capture certain aspects of language history, namely the vertical relations

Waves are bad because

- nobody knows how to reconstruct them
- languages still separate, even if not in split processes

Phylogenetic Networks

Trees are bad because

- they are difficult to reconstruct
- languages do not separate in split processes
- they are boring, since they only capture certain aspects of language history, namely the vertical relations

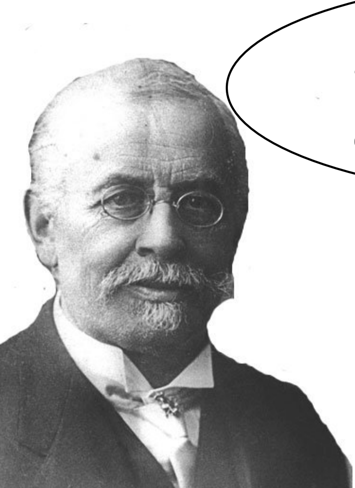
Waves are bad because

- nobody knows how to reconstruct them
- languages still separate, even if not in split processes
- they are boring, since they only capture certain aspects of language history, namely, the horizontal relations

Phylogenetic Networks

Hugo Schuchardt
(1842-1927)

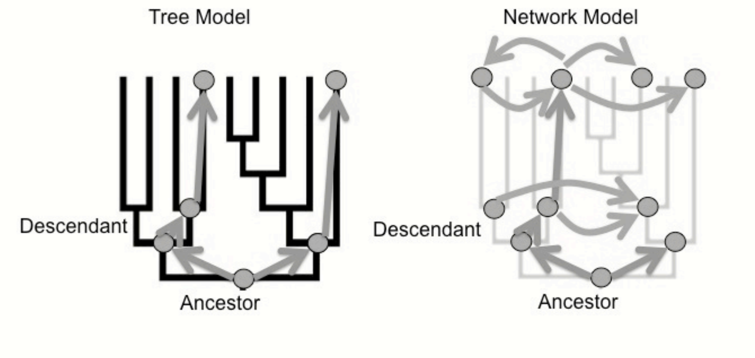
Phylogenetic Networks



We connect the branches and twigs of the tree with countless horizontal lines and it ceases to be a tree (Schuchardt 1870 [1900]: 11)

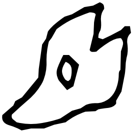
Hugo Schuchardt (1842-1927)

Phylogenetic Networks



魚 ?

Modelling Chinese Dialect History

首 

Data

Data

- The data for this study was taken from the Xiàndài Hànyǔ Fāngyán Yīnkù (Hou 2004).

Data

- The data for this study was taken from the Xiàndài Hànyǔ Fāngyán Yīnkù (Hou 2004).
- It consists of 180 items (“meanings”) translated into 40 contemporary Chinese dialects.

Data

- The data for this study was taken from the Xiàndài Hànyǔ Fāngyán Yīnkù (Hou 2004).
- It consists of 180 items (“meanings”) translated into 40 contemporary Chinese dialects.
- The data is available on a CD in RTF format along with recordings for all dialect entries.

Data

- The data for this study was taken from the Xiàndài Hànyǔ Fāngyán Yīnkù (Hou 2004).
- It consists of 180 items (“meanings”) translated into 40 contemporary Chinese dialects.
- The data is available on a CD in RTF format along with recordings for all dialect entries.
- For this study, the transcriptions in RTF were converted to Unicode.

Data

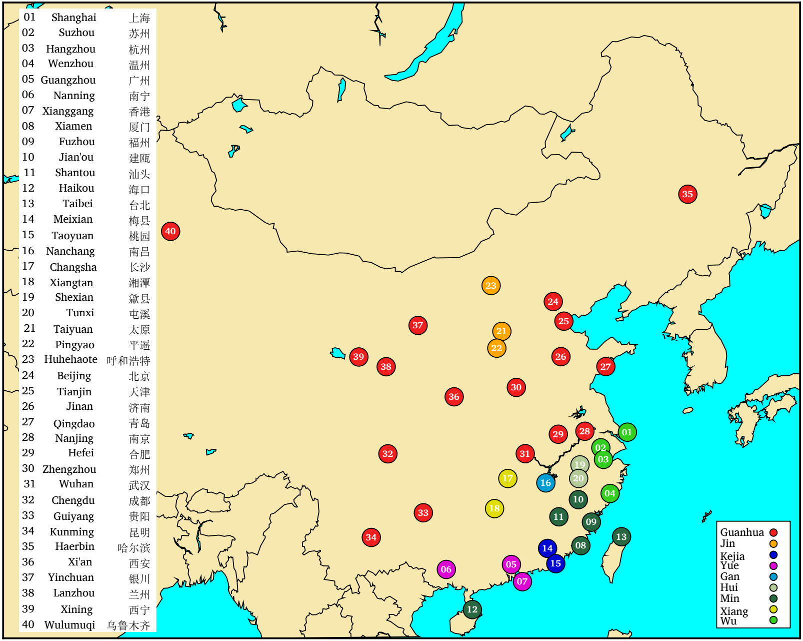
- The data for this study was taken from the Xiàndài Hànyǔ Fāngyán Yīnkù (Hou 2004).
- It consists of 180 items (“meanings”) translated into 40 contemporary Chinese dialects.
- The data is available on a CD in RTF format along with recordings for all dialect entries.
- For this study, the transcriptions in RTF were converted to Unicode.
- Every word was compared with the recordings in order to minimize errors resulting from the extraction process and the original encoding itself.

Data

ITEM 太阳 *tàiyáng* “sun”

Dialect	Pronunciation	Characters	Cognacy
Shanghai	tʰa$^{34\text{-}33}$ɦiã$^{13\text{-}44}$	太阳	1
Shanghai	ȵjɪʔ$^{1\text{-}11}$dɤ$^{13\text{-}23}$	日头	2
Wenzhou	tʰa$^{42\text{-}22}$ji	太阳	1
Wenzhou	ȵi$^{213\text{-}22}$dɤu	日头	2
Guangzhou	jit^{2}tʰɐu$^{21\text{-}35}$	热头	3
Guangzhou	tʰai^{33}jœŋ^{21}	太阳	1
Haikou	zit^{3}hau^{31}	日头	2
Beijing	tʰai^{51}iɑŋ^{1}	太阳	1

Dialect Locations in the Xiàndài Hànyǔ Fāngyán Yīnkù



01	Shanghai	上海
02	Suzhou	苏州
03	Hangzhou	杭州
04	Wenzhou	温州
05	Guangzhou	广州
06	Nanning	南宁
07	Xianggang	香港
08	Xiamen	厦门
09	Fuzhou	福州
10	Jian'ou	建瓯
11	Shantou	汕头
12	Haikou	海口
13	Taibei	台北
14	Meixian	梅县
15	Taoyuan	桃园
16	Nanchang	南昌
17	Changsha	长沙
18	Xiangtan	湘潭
19	Shexian	歙县
20	Tunxi	屯溪
21	Taiyuan	太原
22	Pingyao	平遥
23	Huhehaote	呼和浩特
24	Beijing	北京
25	Tianjin	天津
26	Jinan	济南
27	Qingdao	青岛
28	Nanjing	南京
29	Hefei	合肥
30	Zhengzhou	郑州
31	Wuhan	武汉
32	Chengdu	成都
33	Guiyang	贵阳
34	Kunming	昆明
35	Haerbin	哈尔滨
36	Xi'an	西安
37	Yinchuan	银川
38	Lanzhou	兰州
39	Xining	西宁
40	Wulumuqi	乌鲁木齐

Analysis

Analysis

- The data was analyzed with help of Dagan and Martin's (2008) method for phylogenetic network reconstruction, that was applied to linguistic data before (Nelson-Sathi et al. 2011).

Analysis

- The data was analyzed with help of Dagan and Martin's (2008) method for phylogenetic network reconstruction, that was applied to linguistic data before (Nelson-Sathi et al. 2011).
- Given a binary reference tree reflecting the vertical history of a language family and a list of homologs ("cognates") distributed over the languages, the method reconstructs horizontal relations between the languages and the internal nodes of the tree.

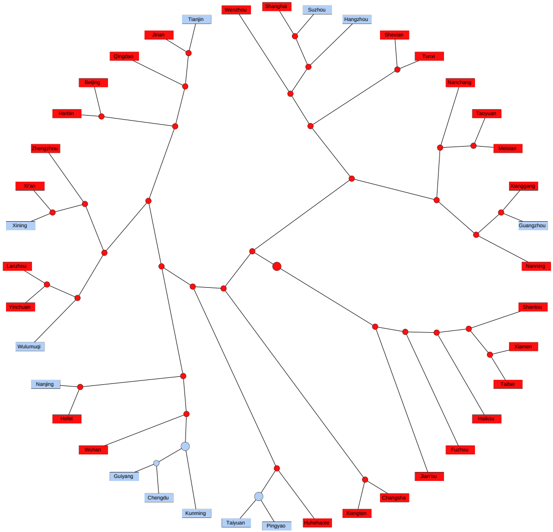
Analysis

- The data was analyzed with help of Dagan and Martin's (2008) method for phylogenetic network reconstruction, that was applied to linguistic data before (Nelson-Sathi et al. 2011).
- Given a binary reference tree reflecting the vertical history of a language family and a list of homologs ("cognates") distributed over the languages, the method reconstructs horizontal relations between the languages and the internal nodes of the tree.
- The reconstruction of horizontal relations is done by seeking specific evolutionary models (loss and gain of characters) that fit the given distribution best.

Analysis

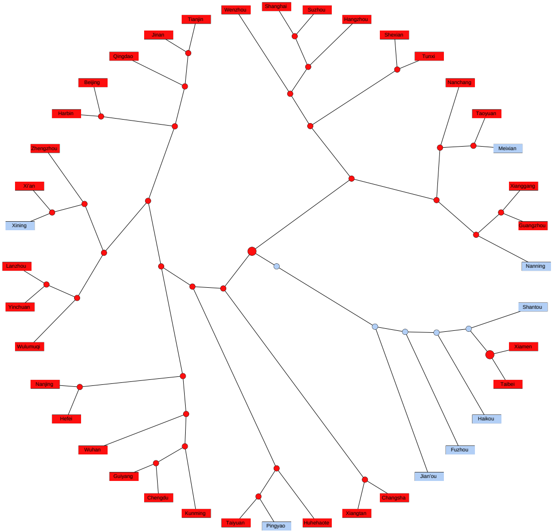
- The data was analyzed with help of Dagan and Martin's (2008) method for phylogenetic network reconstruction, that was applied to linguistic data before (Nelson-Sathi et al. 2011).
- Given a binary reference tree reflecting the vertical history of a language family and a list of homologs ("cognates") distributed over the languages, the method reconstructs horizontal relations between the languages and the internal nodes of the tree.
- The reconstruction of horizontal relations is done by seeking specific evolutionary models (loss and gain of characters) that fit the given distribution best.
- The main criterion by which the fitness of the distributions is evaluated is the "vocabulary size", i.e. the distribution of word forms over a set of meanings. Comparing the vocabulary sizes of different models that infer different amounts of lateral events, the model that comes closest to the vocabulary sizes of the contemporary languages is chosen.

Analysis



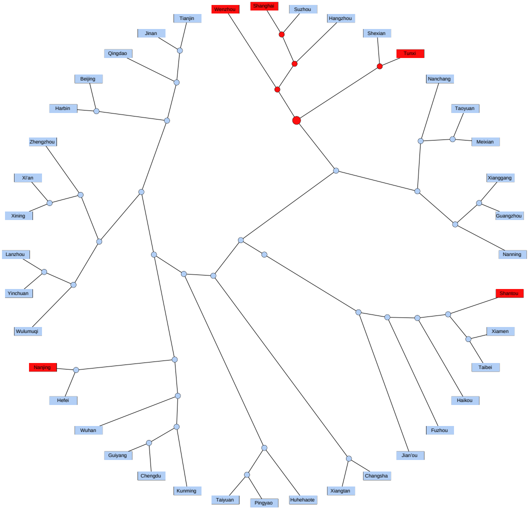
“sun” 日头 *rìtou*

Analysis



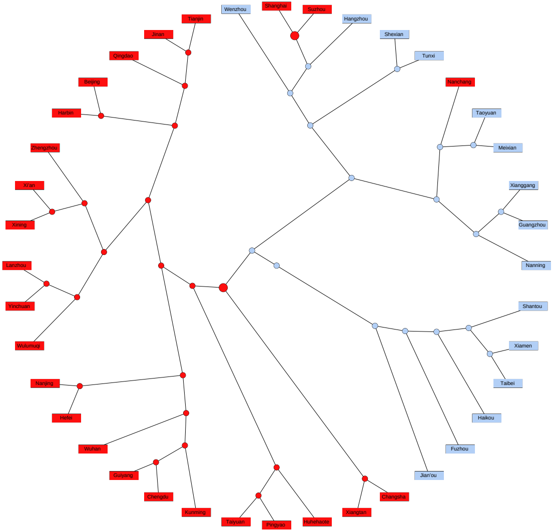
“sun” 太阳 *tàiyáng*

Analysis



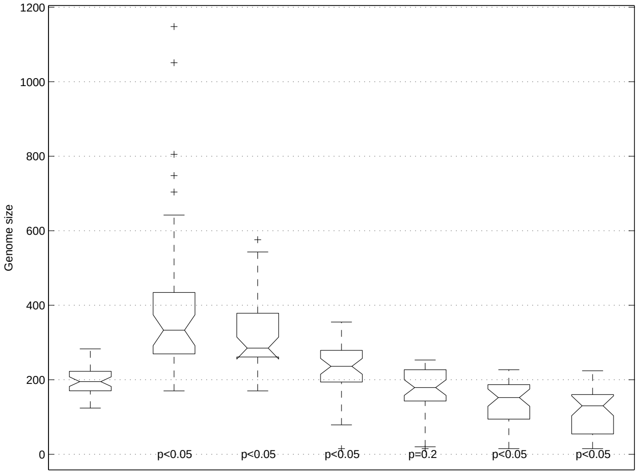
“become sick” 生病 *shēngbìng*

Analysis



“aubergine” 茄子 *qiézi*

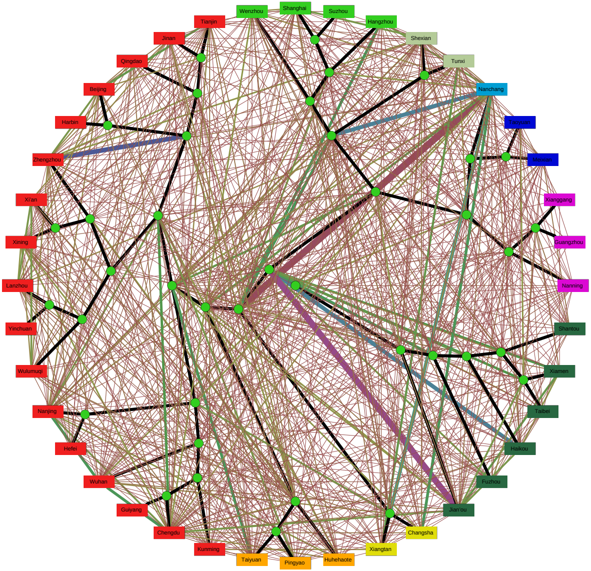
Results


Genome size
0
200
400
600
800
1000
1200
p<0.05
p<0.05
p<0.05
p=0.2
p<0.05
p<0.05

Results

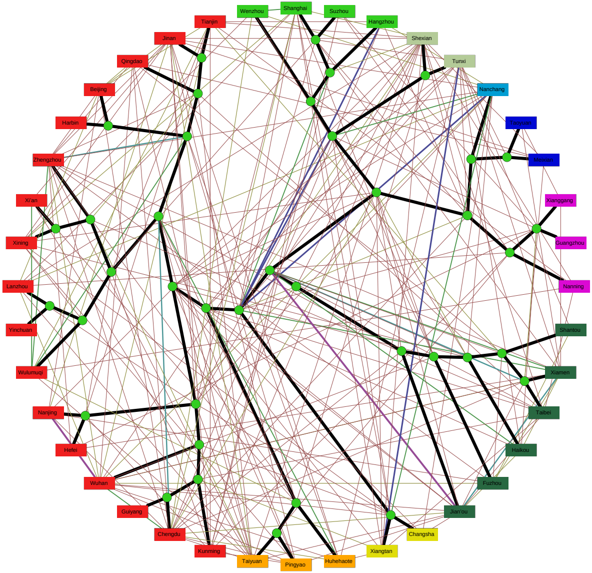
- The BOR3-model fits the distribution best. It allows up to three lateral connections per homolog.
- Out of 1152 homologs distributed over the Chinese dialects, 264 are monophyletic, 328 require one, 355 two, and 177 three lateral links in order to explain the distribution neatly.
- This corresponds to a borrowing rate of 0.5286 borrowing events per homolog per lifetime.
- For 78 percent of all homologs in the dataset the method reconstructs lateral links and therefore suggests that these have been involved in borrowing events during their history.
- Suprisingly, the 48 homologs that correspond to basic vocabulary concepts in the dataset do not show significant differences in their borrowing rates compared to the non-basic items.

Results: General Results



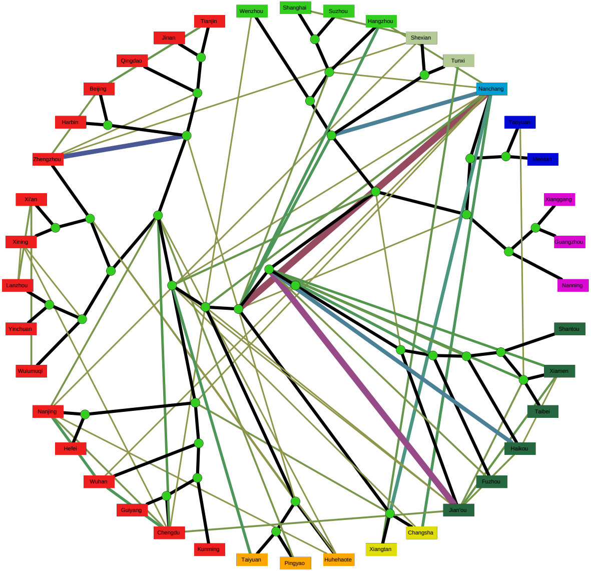
Whole Dataset

Results: General Results



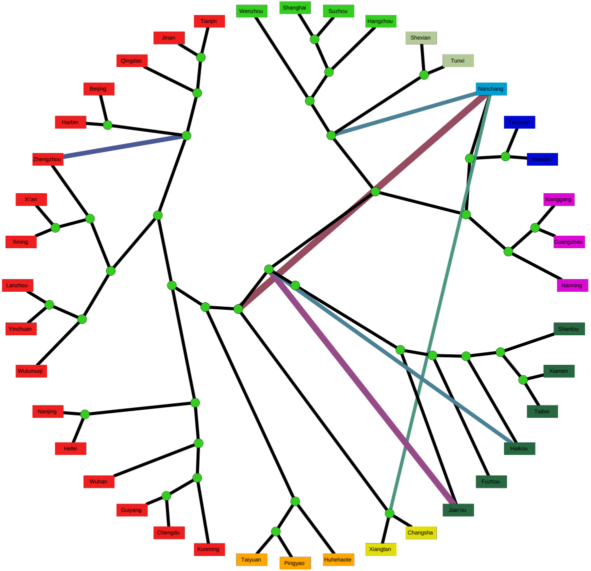
Swadesh Subset

Results: General Results



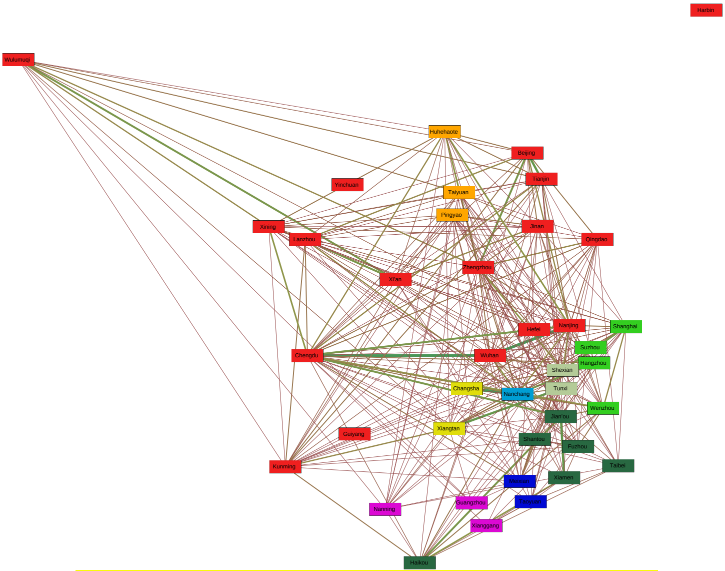
Whole Dataset (Cutoff 5)

Results: General Results



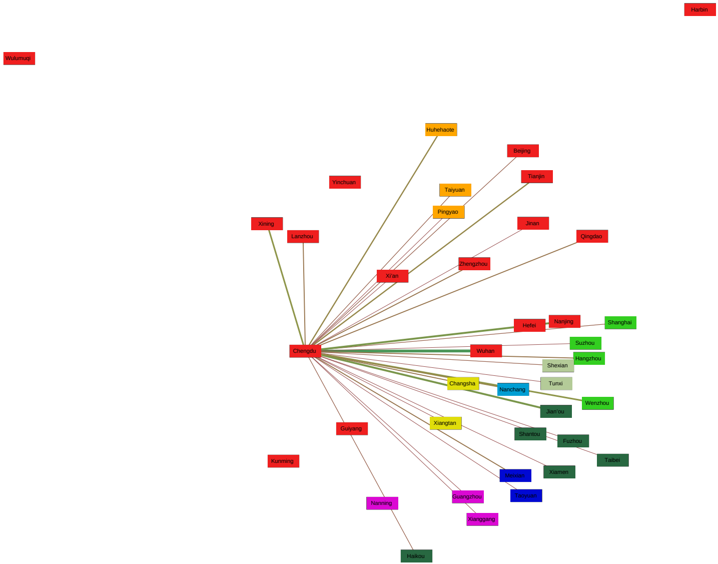
Whole Dataset (Cutoff 10)

Results: Chengdu



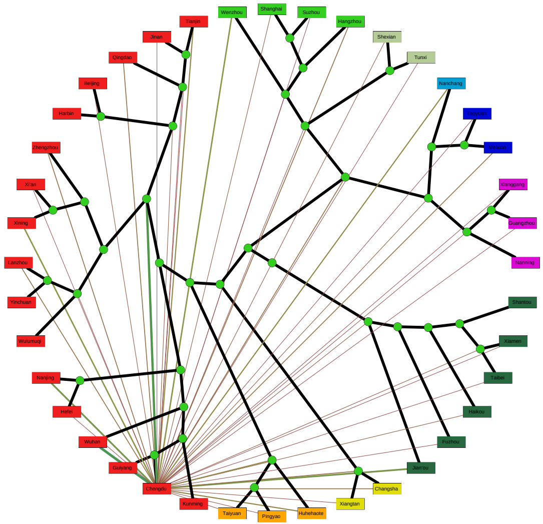
Contemporary Links Mapped to Coordinates

Results: Chengdu



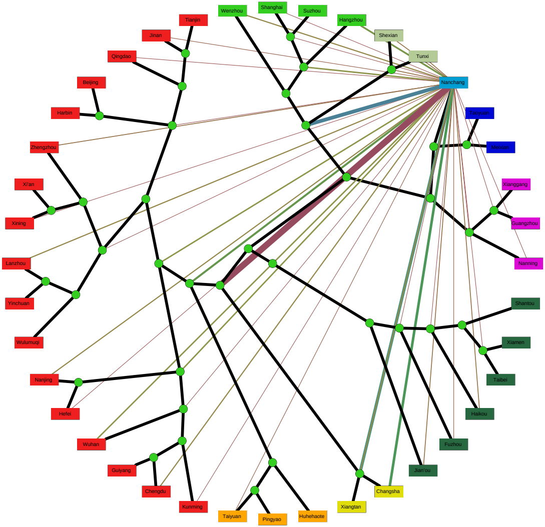
Contemporary Links of Chengdu

Results: Chengdu



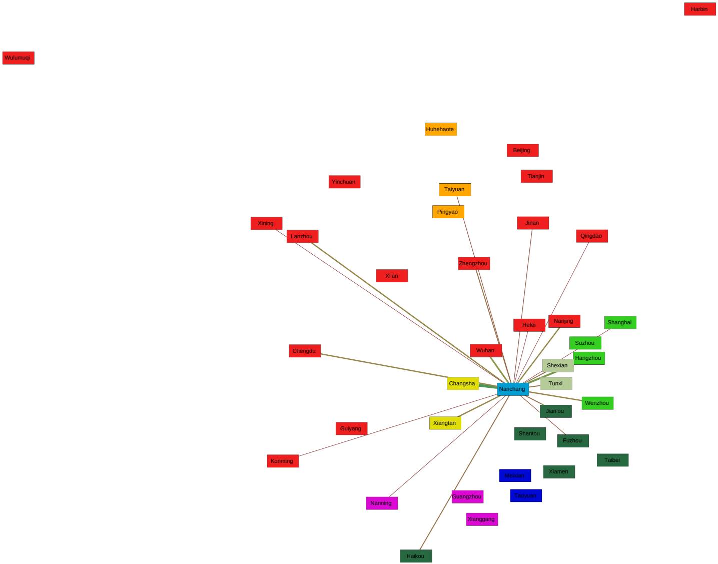
Links of Chengdu

Results: Nanchang



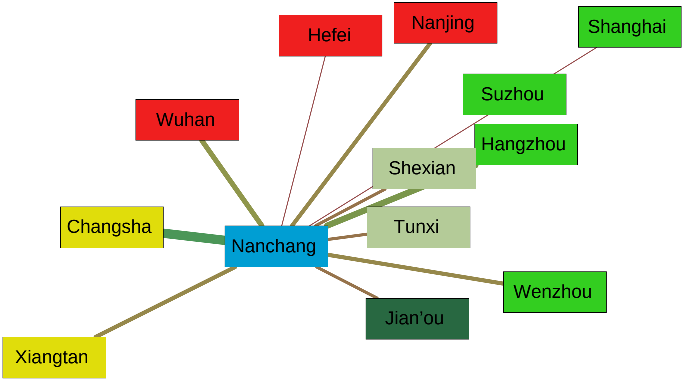
Links of Nanchang

Results: Nanchang



Contemporary Links of Nanchang

Results: Nanchang



Links between Nanchang and its Neighbors

Concluding Remarks

- Phylogenetic networks look nice.
- Phylogenetic networks are – if properly reconstructed – a valid alternative to both the tree and the wave model.
- We need to test the method by Dagan and Martin (2008) on more data and in more detail in order to be able to give an account on its full potential and its limits.

Concluding Remarks

谢谢大家！

Concluding Remarks

Thank you!