

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



语言



язык

Languages



language



språk

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	tʂəŋ ⁵⁵ tsai ⁵³	naɛ ⁵¹	tʂəŋ ⁵⁵ luən ⁵¹
Hakka Chinese	1	i <u>u</u> ³³	it ⁵⁵	pai ³³ a ¹¹	pet ³³ f <u>uŋ³³</u>	tʰuŋ ¹¹	nit ¹¹ tʰeu ¹¹	hɔk ³³	e ⁵³	au ⁵⁵
Shanghai Chinese	1	f <u>i</u> ²²	tʰɑ ⁵⁵	ts <u>ɿ</u> ²¹	po? ³ f <u>oŋ⁴⁴</u>	ta? ⁵	tʰa ³³ f <u>ia</u> ⁴⁴	ts <u>ɿ</u> ³³ h <u>ɔ</u> ⁴⁴	lə? ¹ l <u>e</u> ²³ tsa ⁵³	
Beijing Chinese	2	seɪ ³⁵		də ⁵⁵		pən ³⁵ lɪŋ ²¹	ta ⁵¹			
Hakka Chinese	2	man ³³	jin ¹¹		kʷɔ ⁵⁵	vɔi ⁵³				
Shanghai Chinese	2	sa ³³	jin ⁵⁵	fiə? ²¹		pəŋ ³³ z <u>ɿ</u> ⁴⁴	du ¹³			
Norwegian	1	nu:rauvin'ɳ	ɔ	su:lɳ					krɑŋlət	ɔm
Swedish	1	nu:qanvindən	ɔ	su:lən		tyistadə	əgɔŋne			ɔm
Danish	1	noððnven'ɳ	ʌ	so:l ² n	kʰʌm	enøŋan	i sðøið ²			ʌm ²
Norwegian	2	vem	a	dem	sŋ	va:	ʈɻ	stæfkæstə		
Swedish	2	vem	av	dɔm	sɔm	va		sta:kast		
Danish	2	vem ²	a	b̥m	d̥	va	d̥n	sdæ:gøsd̥e		

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

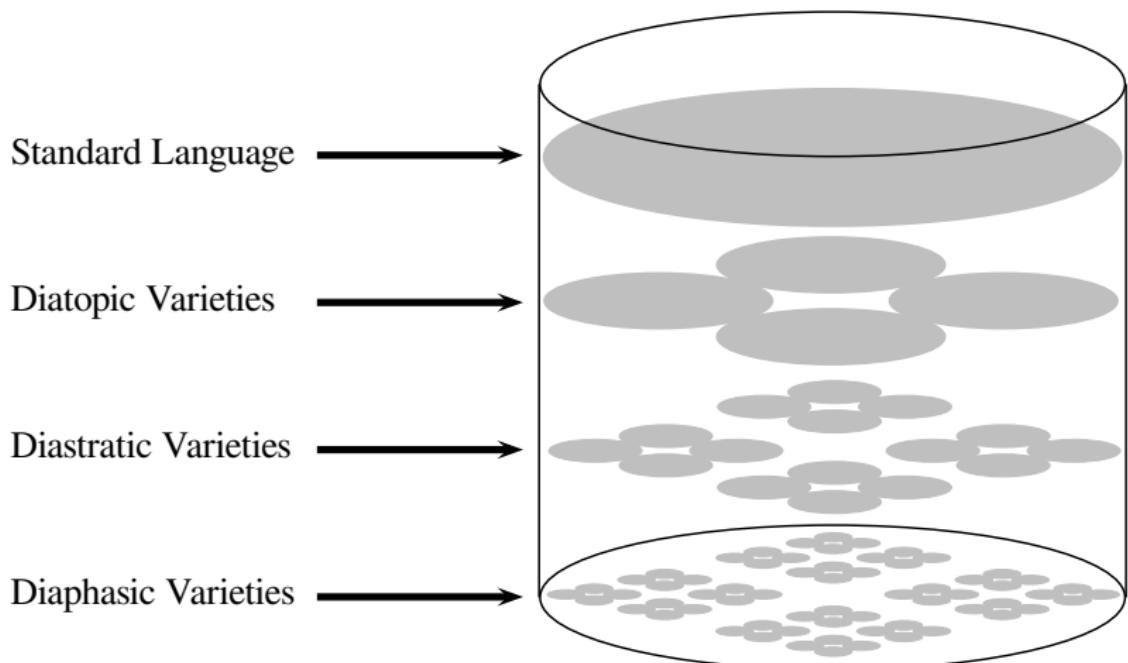
Languages are complex aggregates of different linguistic systems that 'coexist and influence each other' (Coseriu 1973: 40, my translation).

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



Change



Change



expected

Mandarin

[ma₅₅po₂₁lou]

Change



expected

Mandarin

[ma₅₅po₂₁lou]

attested

Mandarin

[wan₅₁paw₂₁lu₅₁]

Change



expected

Mandarin

[ma₅₅po₂₁lou]

attested

Mandarin

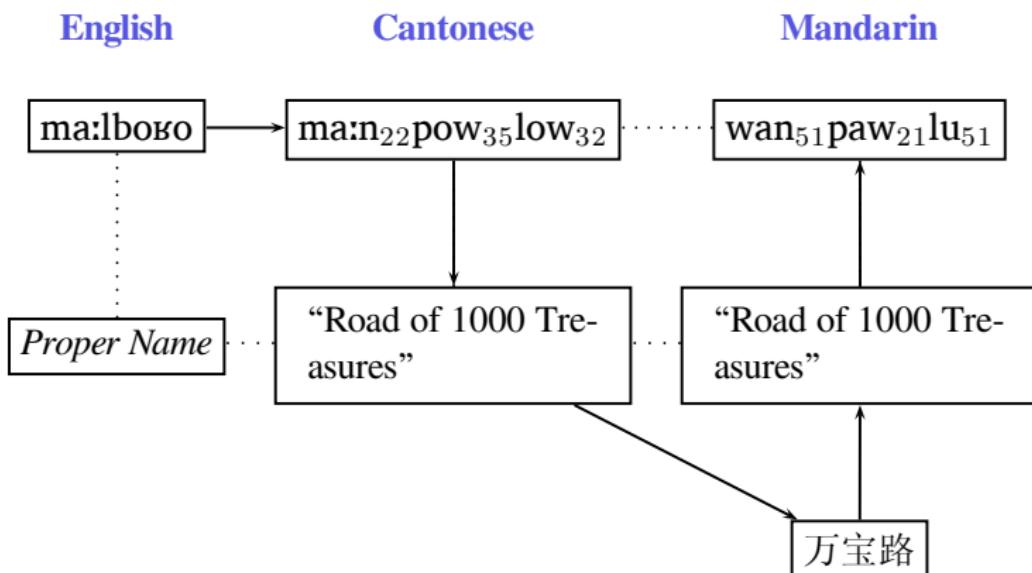
[wan₅₁paw₂₁lu₅₁]

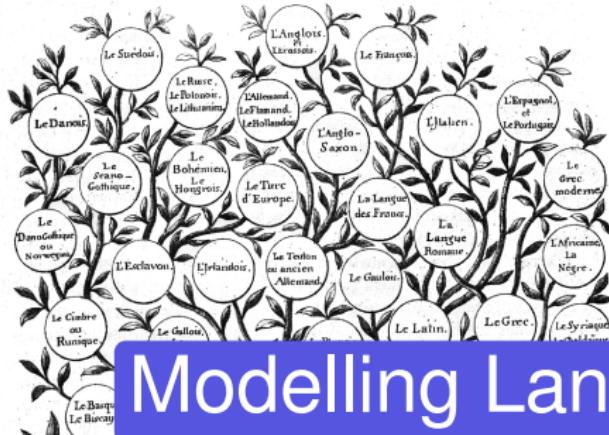
explanation

Cantonese

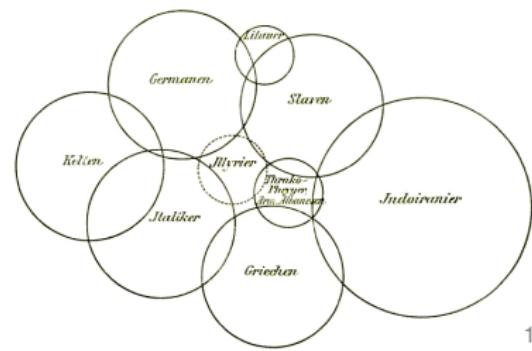
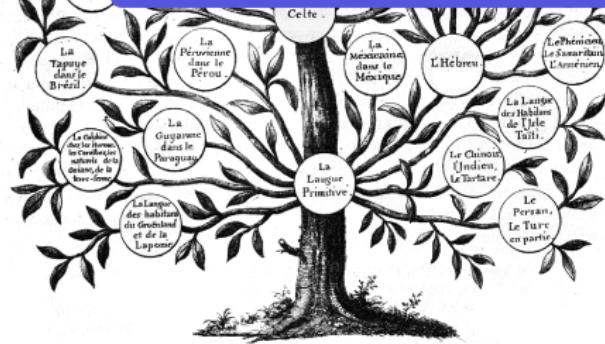
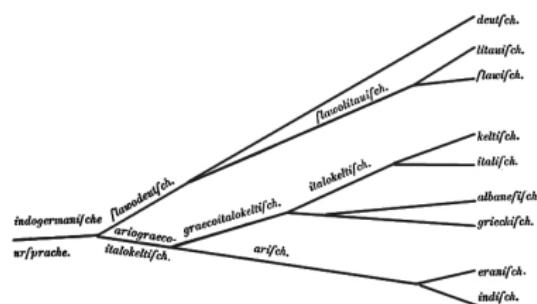
[ma:n₂₂pow₃₅low₃₂]

Change





Modelling Language History



Dendrophilia

August Schleicher
(1821-1868)



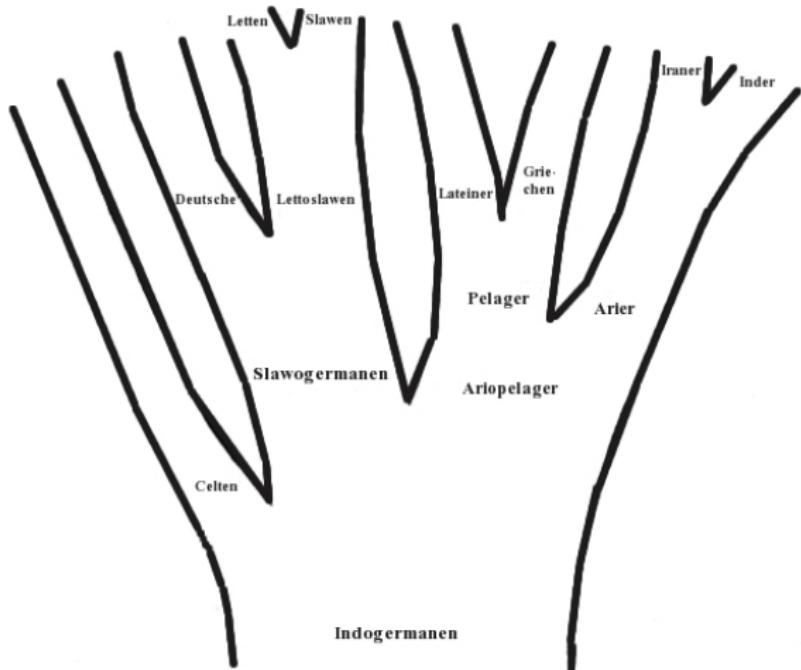
Dendrophilia

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)

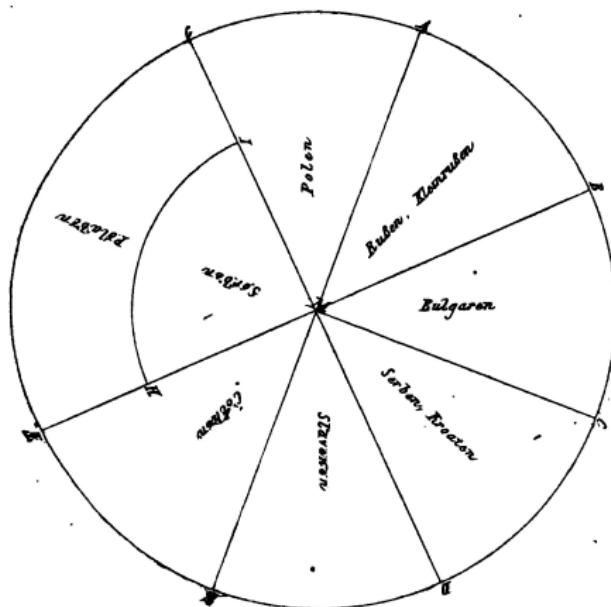
Dendrophobia



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

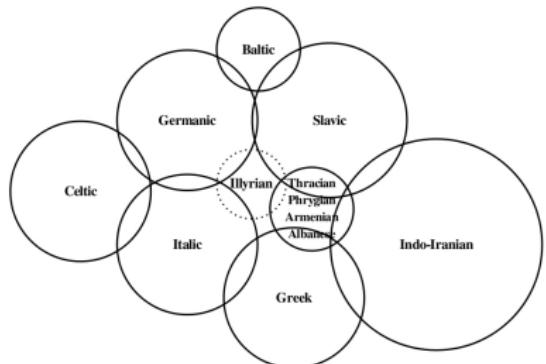


Schmidt (1875)

Dendrophobia



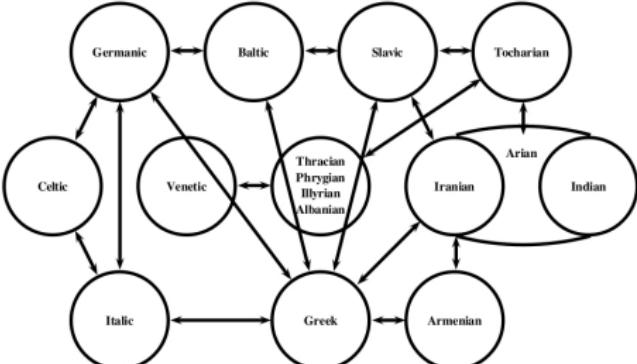
Meillet (1908)



Hirt (1905)



Bloomfield (1933)



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

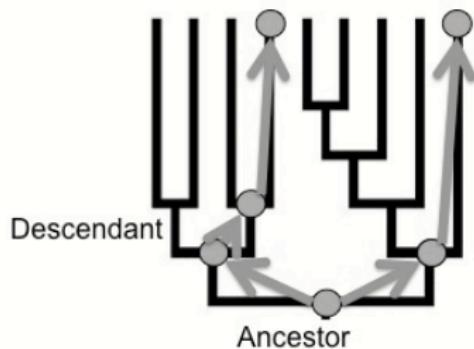


*We connect the branches and twigs
of the tree with countless horizon-
tal lines and it ceases to be a tree*
(Schuchardt 1870 [1900]: 11)

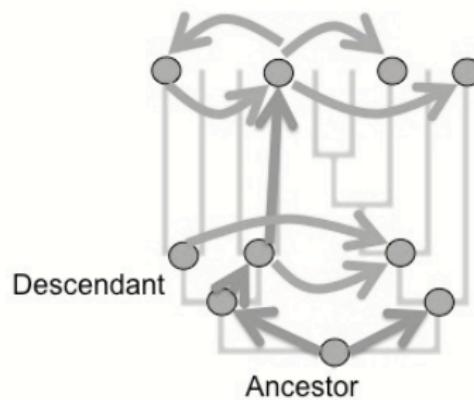
Hugo Schuchardt
(1842-1927)

Phylogenetic Networks

Tree Model



Network Model



魚 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

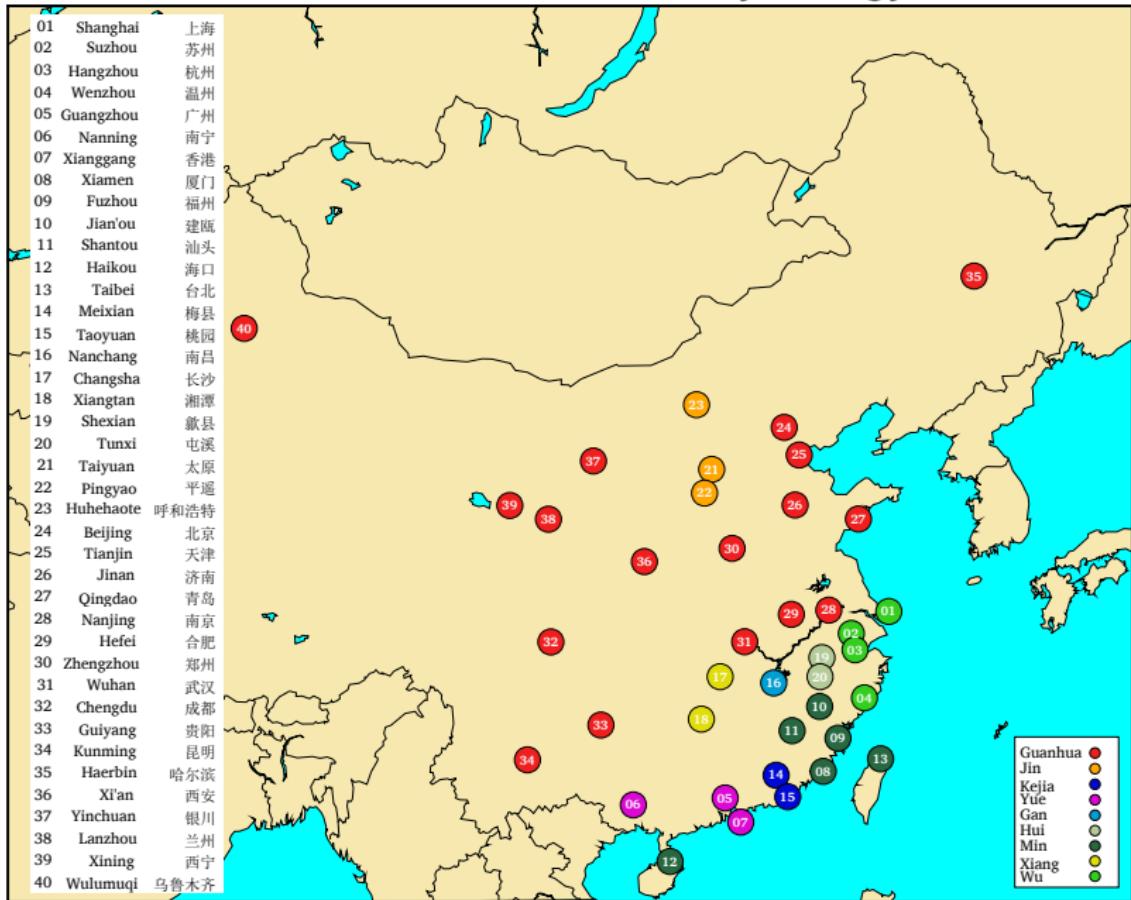
- The data for this study was taken from the *Xiàndài Hànyǔ Fāngyán Yīnkù* (Hou 2004).
- It consists of 180 items (“meanings”) translated into 40 contemporary Chinese dialects.
- The data is available on a CD in RTF format along with recordings for all dialect entries.
- For this study, the transcriptions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

Dialect	Pronunciation	Characters	Cognacy
Shanghai	tʰa ³⁴⁻³³ fiā ¹³⁻⁴⁴	太阳	1
Shanghai	nji? ¹⁻¹¹ dɤ ¹³⁻²³	日头	2
Wenzhou	tʰa ⁴²⁻²² ji	太阳	1
Wenzhou	n̩i ²¹³⁻²² dɤu	日头	2
Guangzhou	jit ² tʰəu ²¹⁻³⁵	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



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.

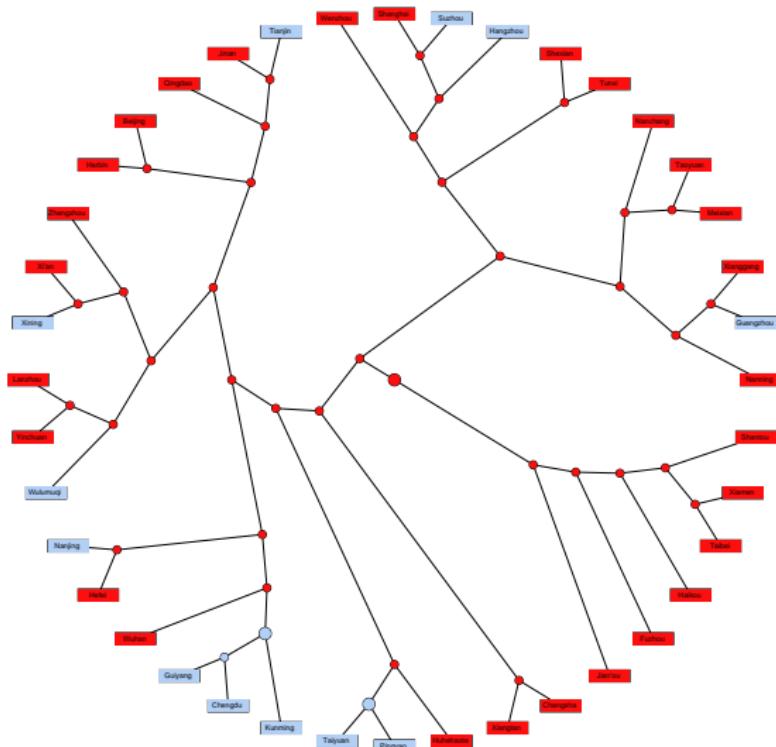
Analysis

- 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

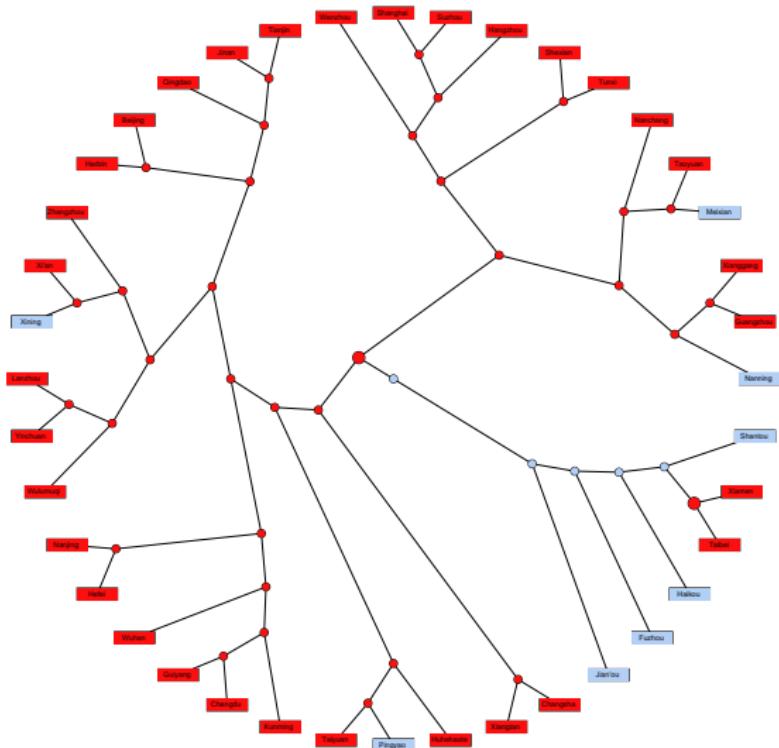
- The data was analyzed with help of Dagan and Martin's (2008) method for phylogenetic network reconstruction, that was applied to linguistic data before (Nelson-Sathi et al. 2011).
- Given a binary reference tree reflecting the vertical history of a language family and a list of homologs ("cognates")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



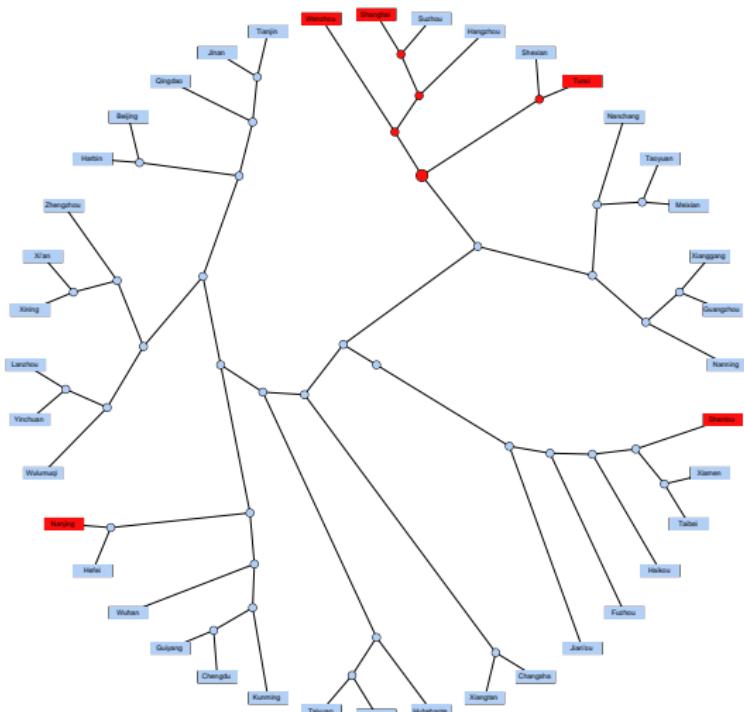
“sun” 日头 *ritou*

Analysis



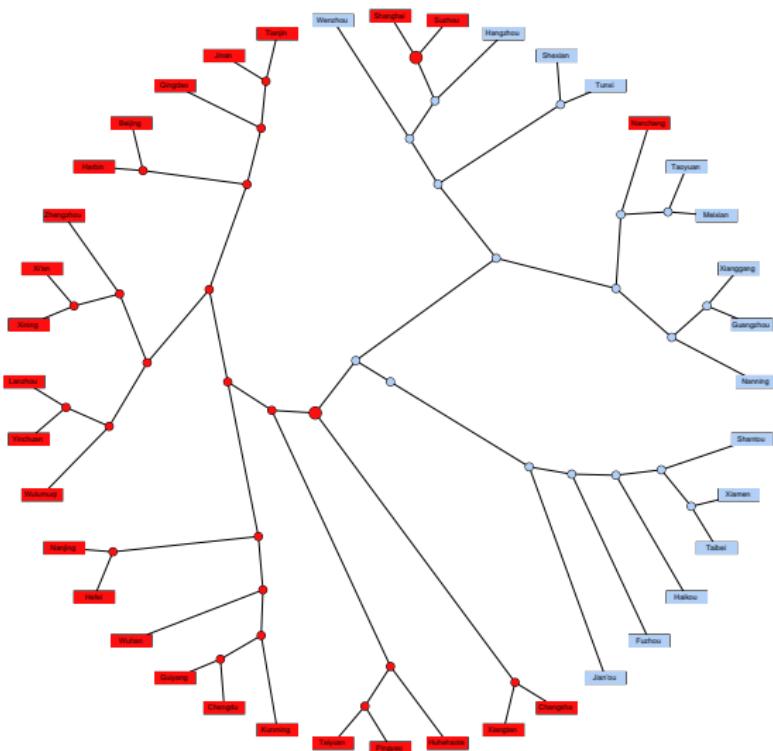
“sun” 太阳 *tài yáng*

Analysis



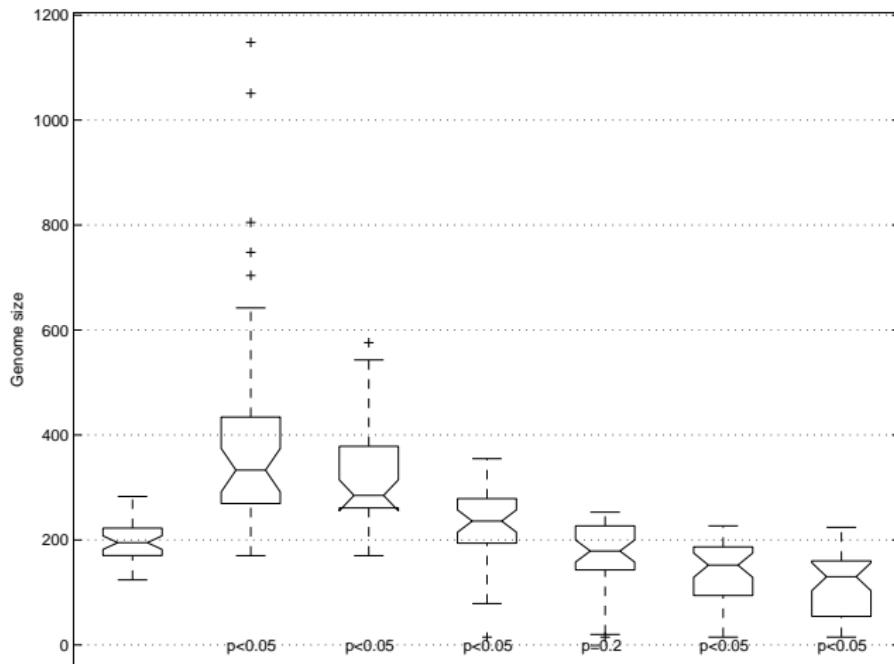
“become sick” 生病 *shēngbìng*

Analysis



“aubergine” 茄子 *qiézi*

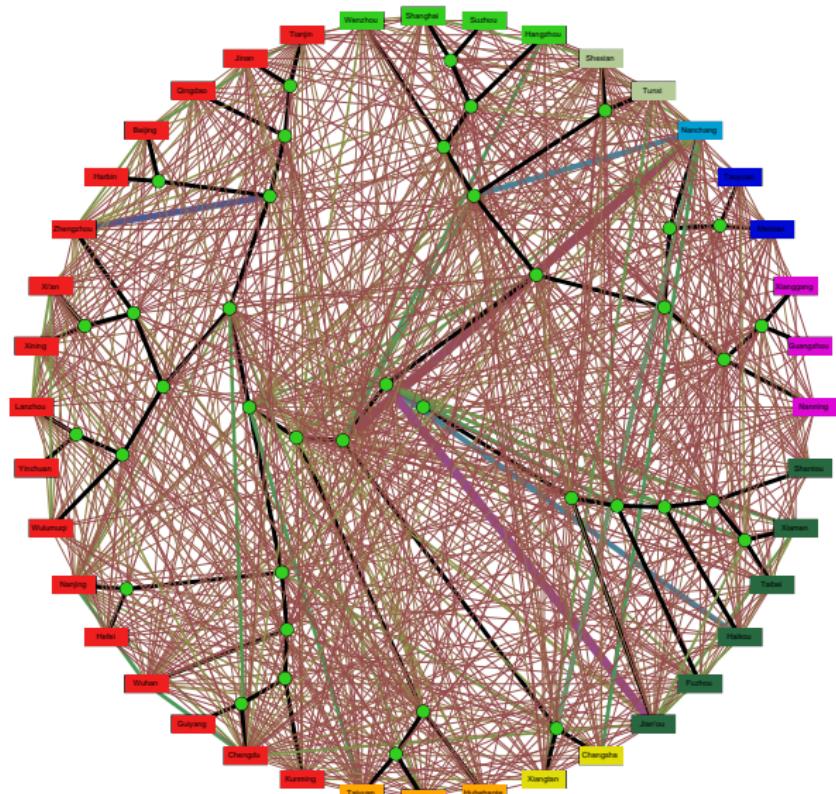
Results



Results

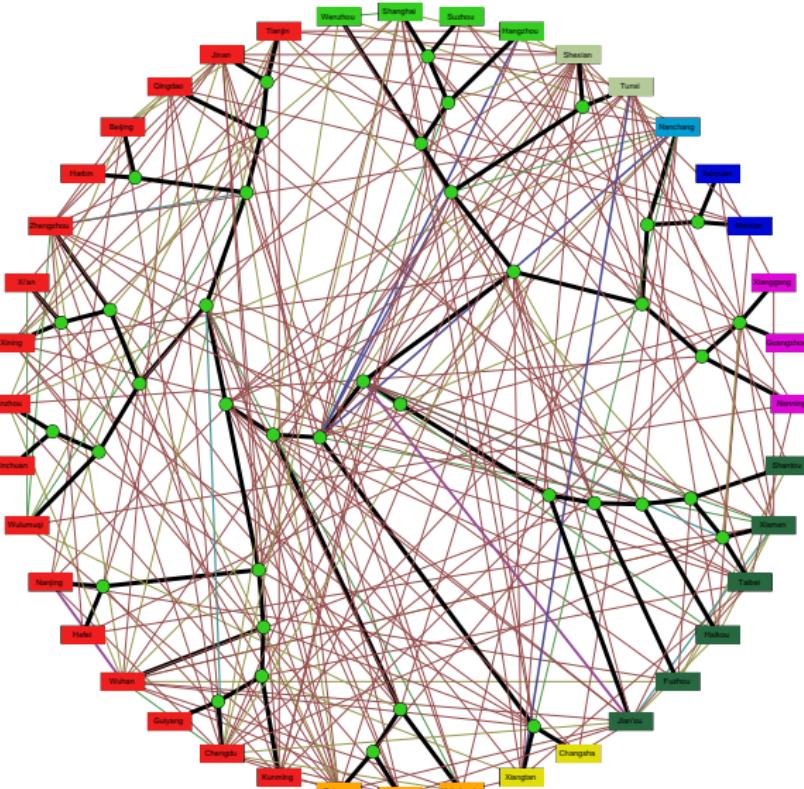
- The BOR3-model fits the distribution best. It allows up to three lateral connections per homolog.
- Out of 1152 homologs distributed over the Chinese dialects, 264 are monophyletic, 328 require one, 355 two, and 177 three lateral links in order to explain the distribution neatly.
- This corresponds 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.
- Surprisingly, 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

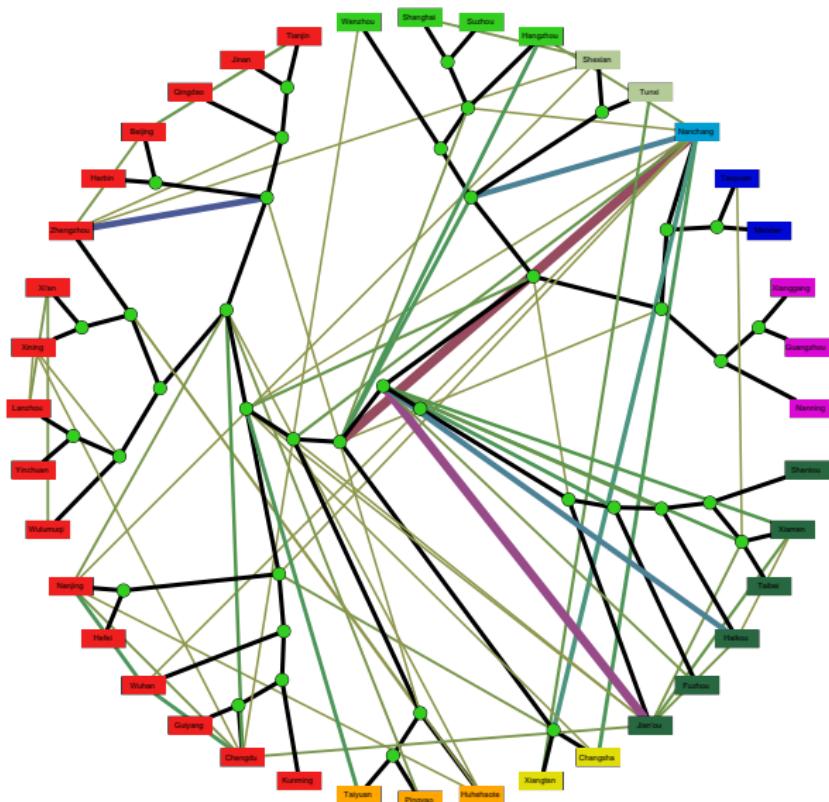


Whole Dataset

Results: General Results

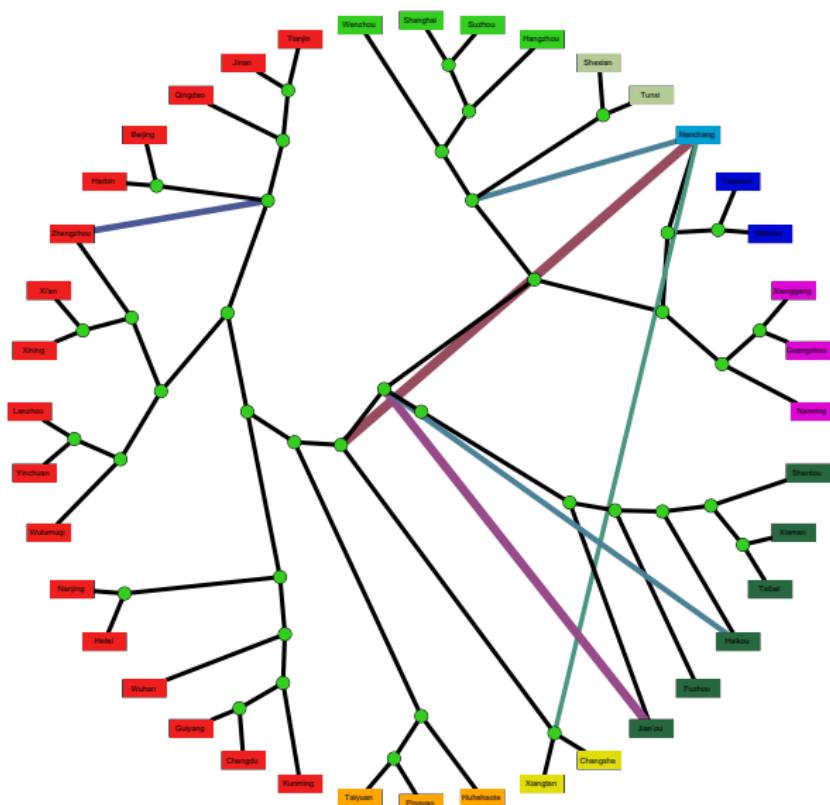


Results: General Results



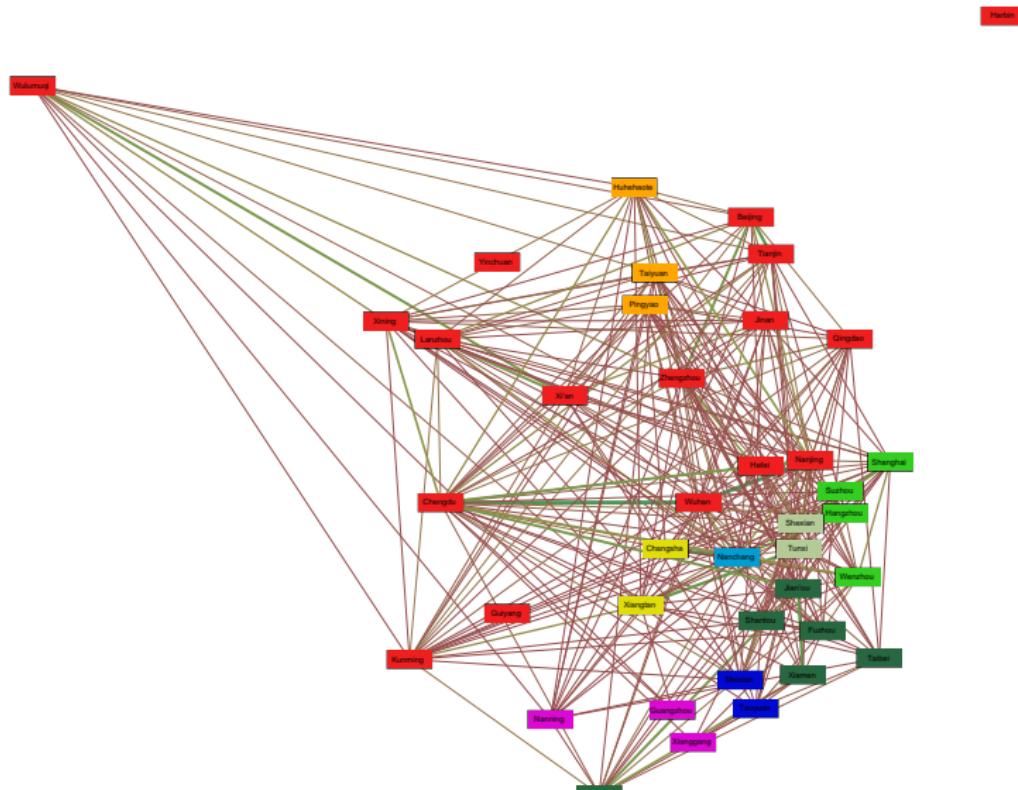
Whole Dataset (Cutoff 5)

Results: General Results



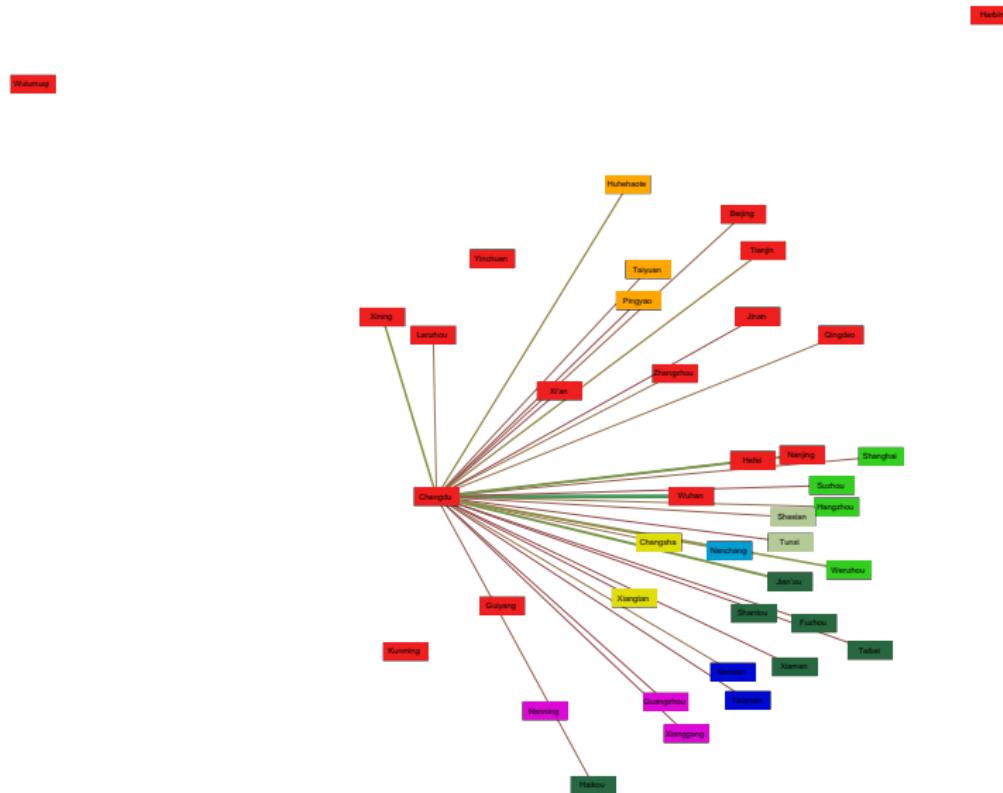
Whole Dataset (Cutoff 10)

Results: Chengdu



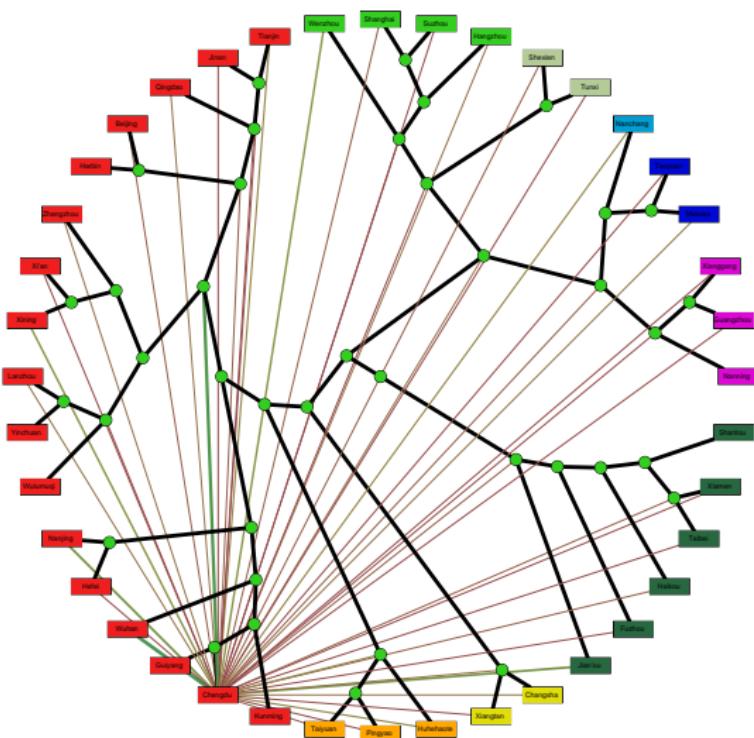
Contemporary Links Mapped to Coordinates

Results: Chengdu

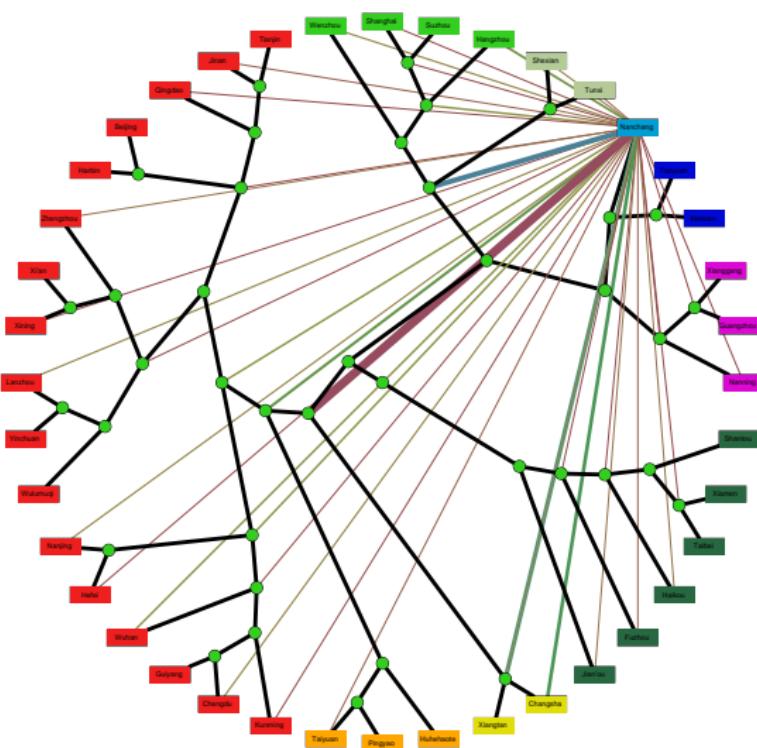


Contemporary Links of Chengdu

Results: Chengdu

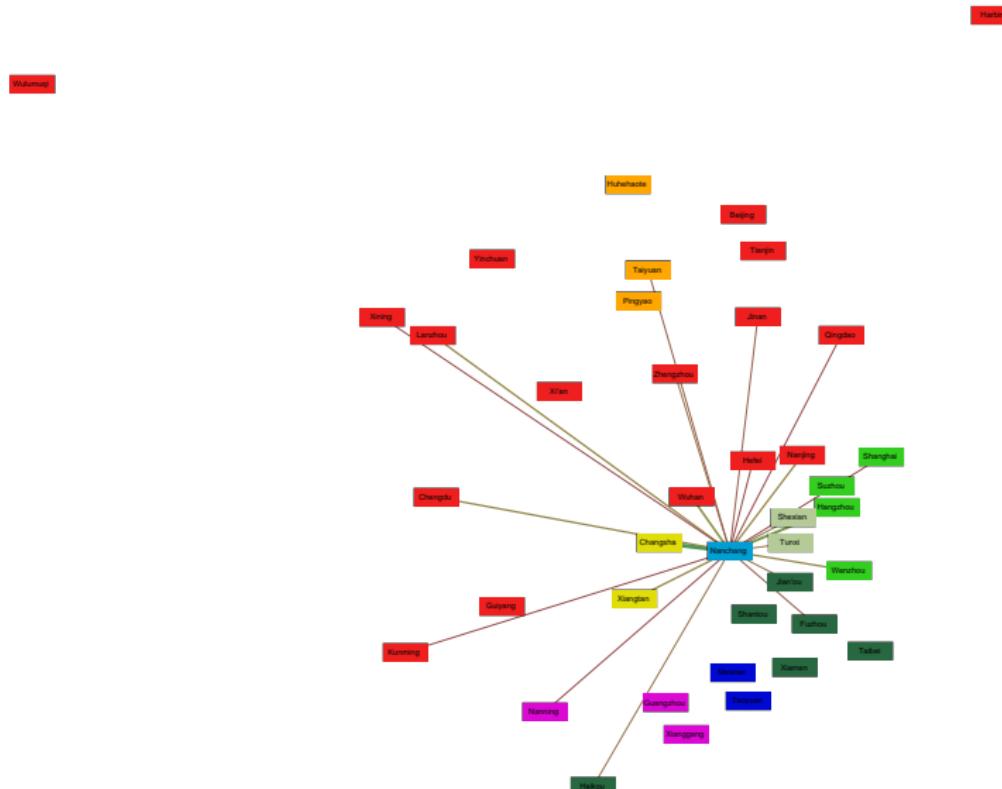


Results: Nanchang



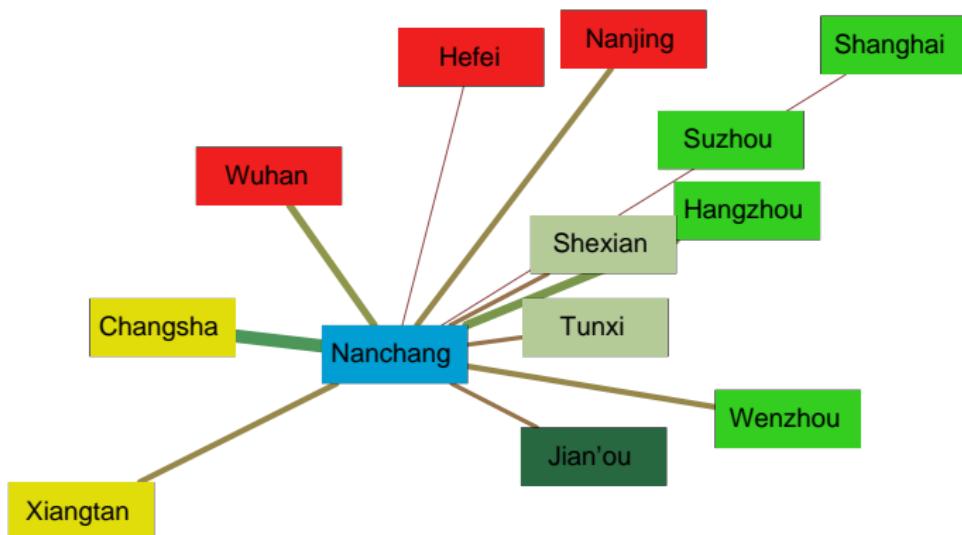
Links of Nanchang

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!