Network Approaches Reveal the Complexity of Chinese Dialect History

Johann-Mattis List*

*Research Center Deutscher Sprachatlas
Philipps-University Marburg

2013/09/26


语言
язык
Languages
language
språk

Languages and Dialects

Norwegian, Danish, and Swedish are different languages.

Běijīng-Chinese, Shànghǎi-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}	t͡ʂəŋ^{55}	tsai^{53}	naɚ^{51}	t͡ʂəŋ^{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} tsɿ^{21}	poʔ^{3}foŋ^{44}	taʔ^{5}	tʰa^{33}ɦiã^{44}	tsəŋ^{33}	hɔ^{44}		ləʔ^{1}lə^{23}tsa^{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	nuːɾɑʊinˑn̩	ɔ	suːlɳ̩					kɾɑŋlət	ɔm	
Swedish	1	nuːɖanvɪndən	ɔ	suːlən		tʏ̥ɪstadə	ən gɔŋ			ɔm	
Danish	1	noˌʌnvenˀn̩	ʌ	soːlˀn	kʰʌm		eŋg̊ɑŋ	i sɖʁiðˀ		ʌmˀ	

Norwegian	2	ʊem	ɑ	dem	sm̩	ʊɑː	ɖɳ̩	stæɾ̥kəstə			
Swedish	2	vɛm	ɑv	dɔm	sɔm	vɑ		staɹkast			
Danish	2	vɛmˀ	a	b̥m̩	ɖ	vɑ	ɖn̩	sɖæˌg̊əsɖə			

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

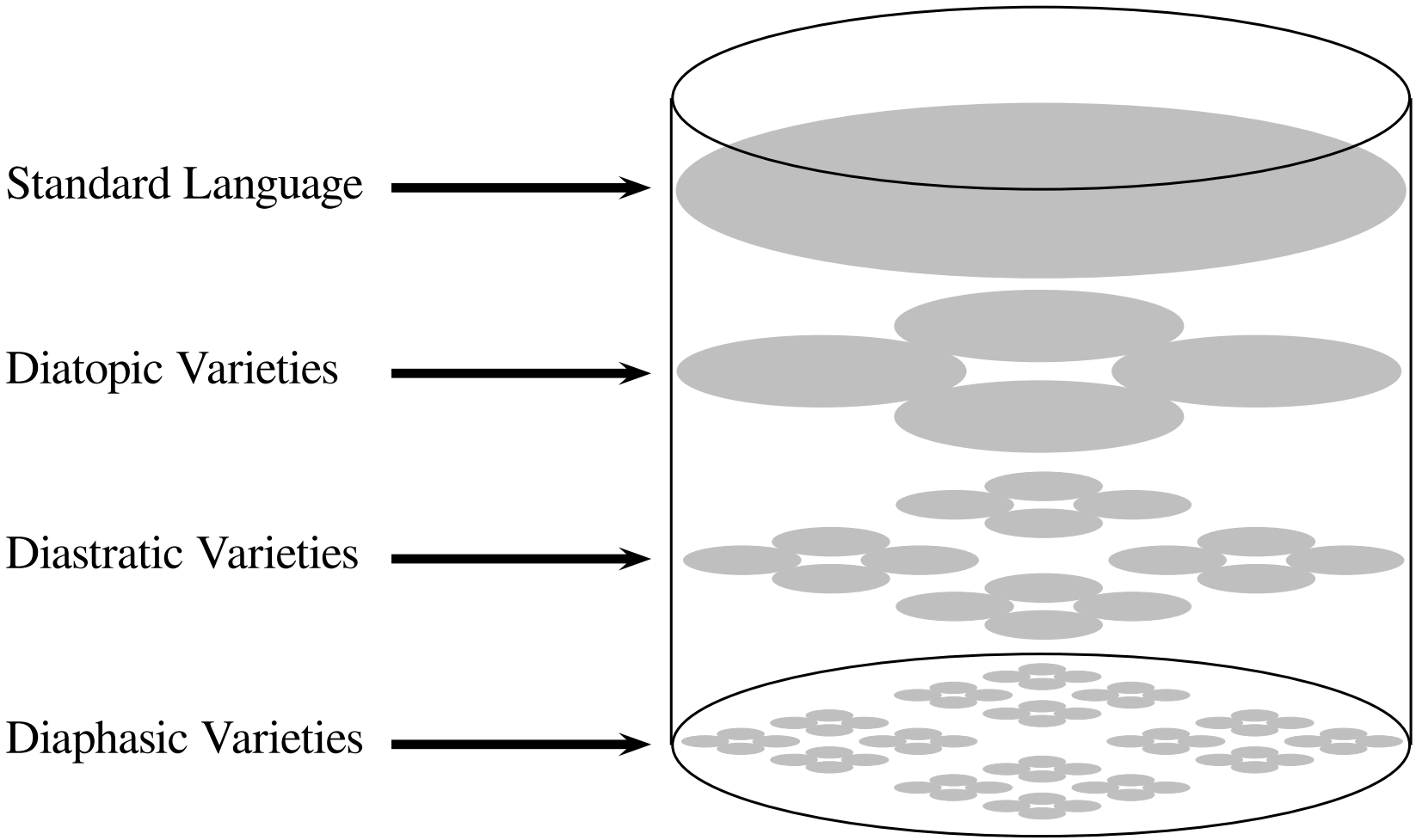
Languages are complex aggregates of different linguistic systems that 'coexist and influence each other' (Coseriu 1973: 40, my translation).

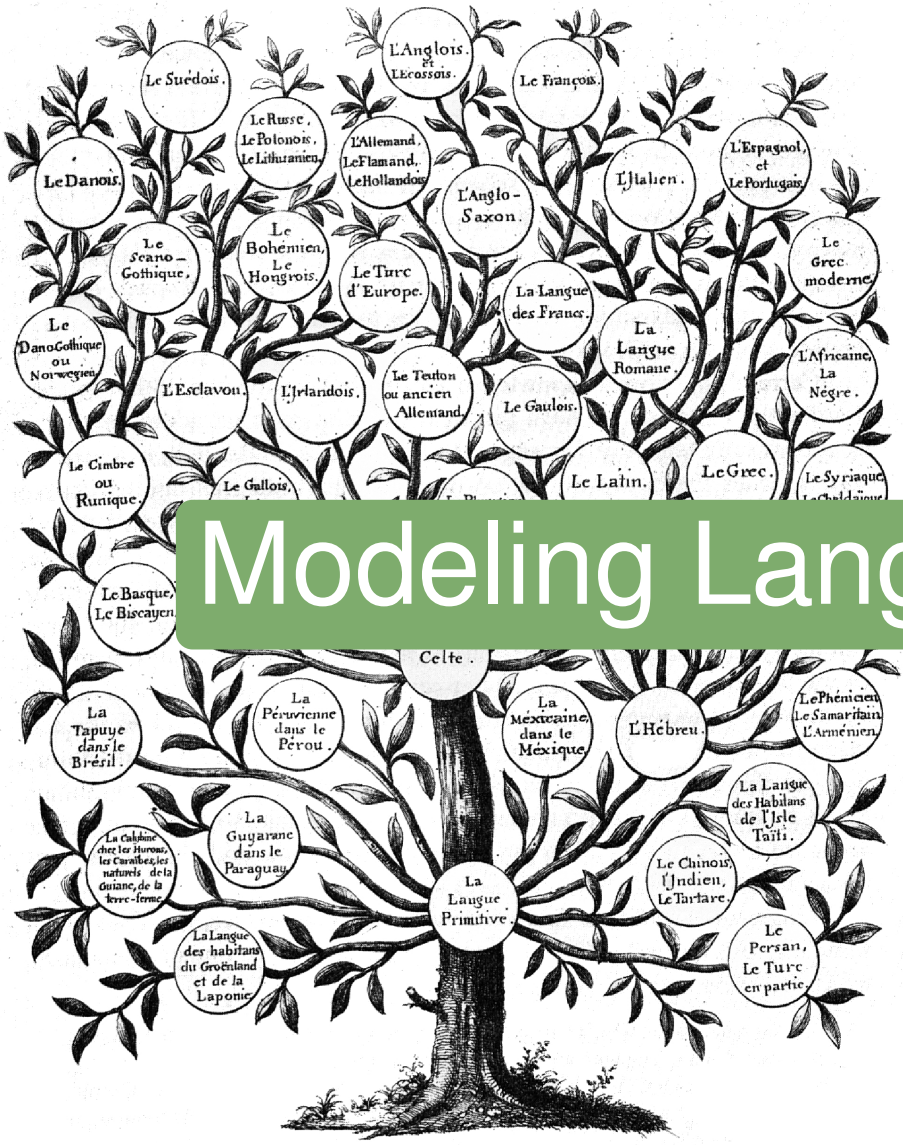
Language as a Diasystem

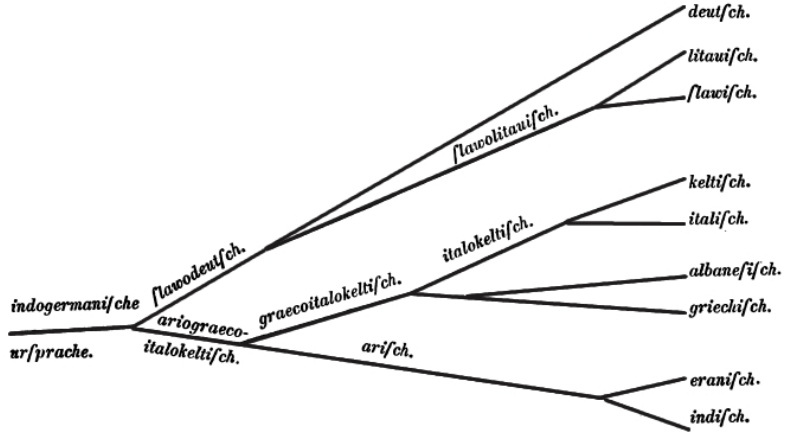
Languages are complex aggregates of different linguistic systems that 'coexist and influence each other' (Coseriu 1973: 40, my translation).

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

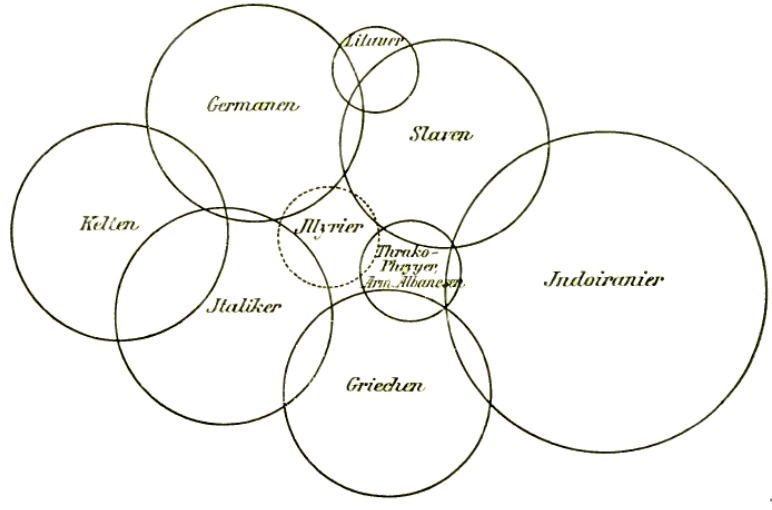
Language as a Diasystem







Modeling Language History



Dendrophilia

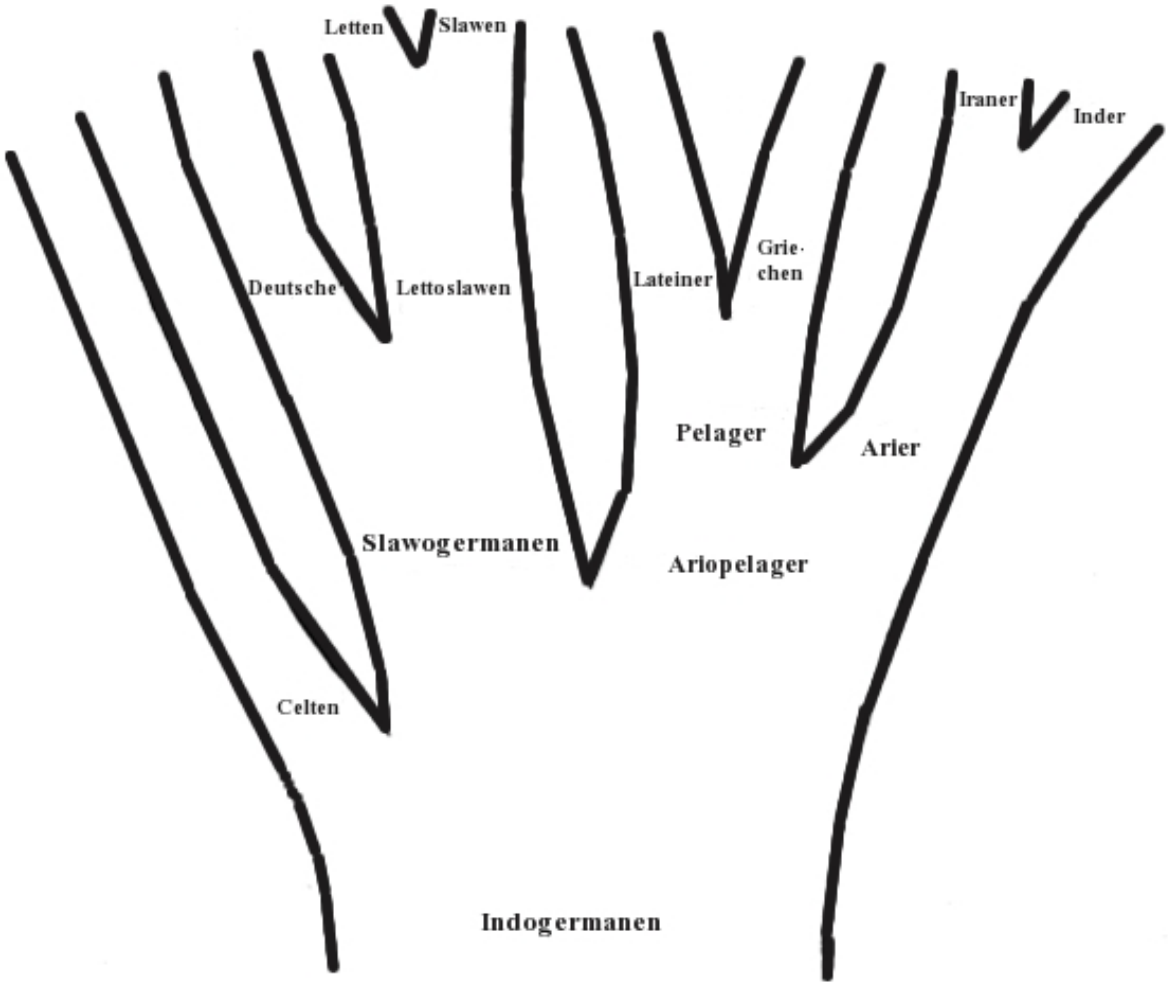
August Schleicher
(1821-1868)

Dendrophilia

These assumptions that logically follow from the results of our re- search 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)

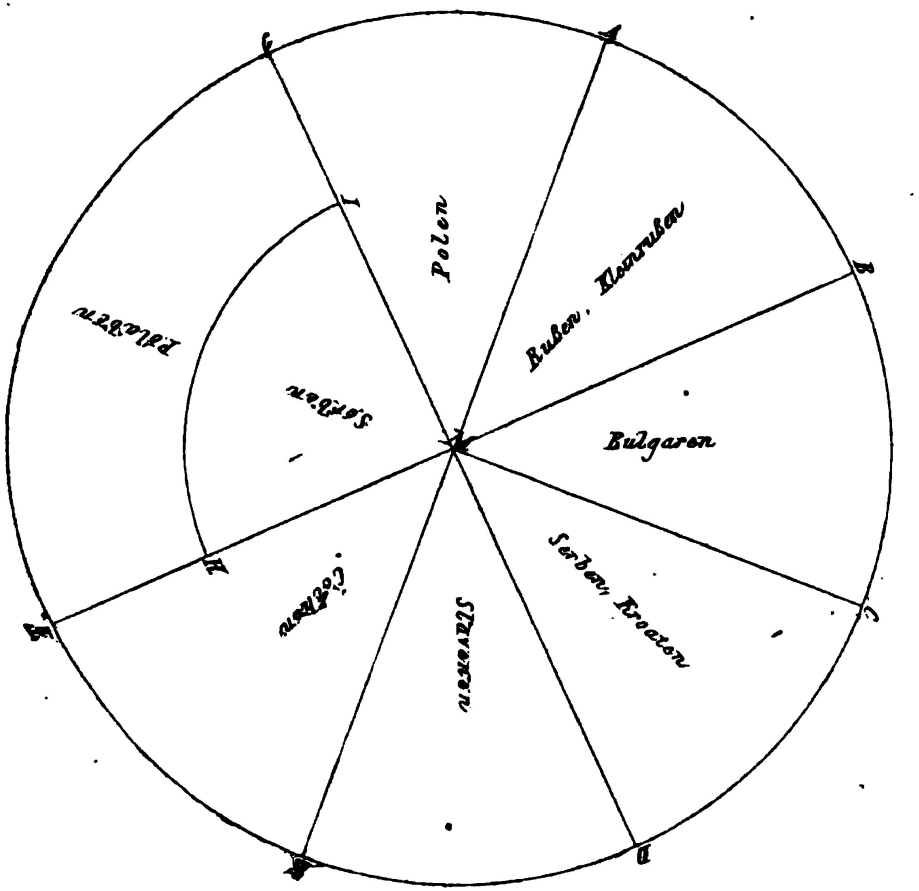
Johannes Schmidt
(1843-1901)

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)

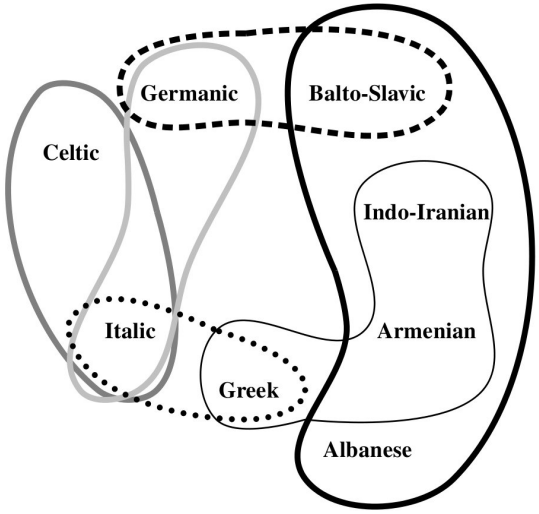
Dendrophobia



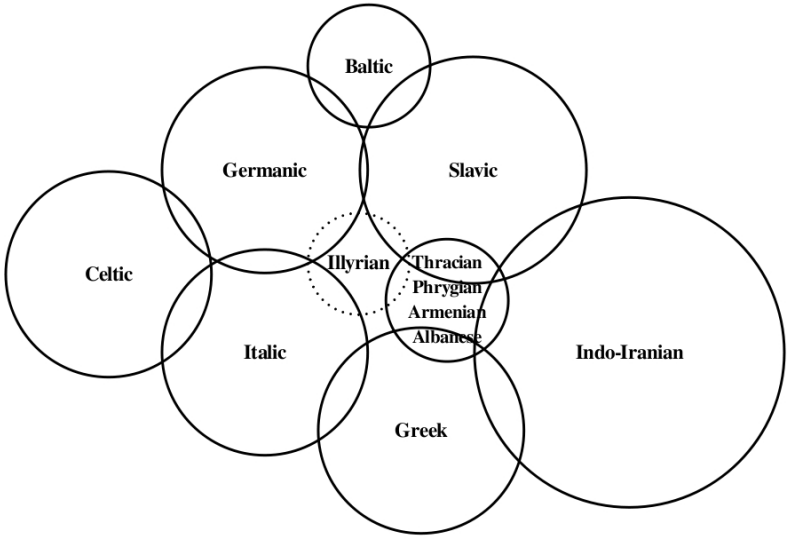
Schmidt (1875)

Dendrophobia

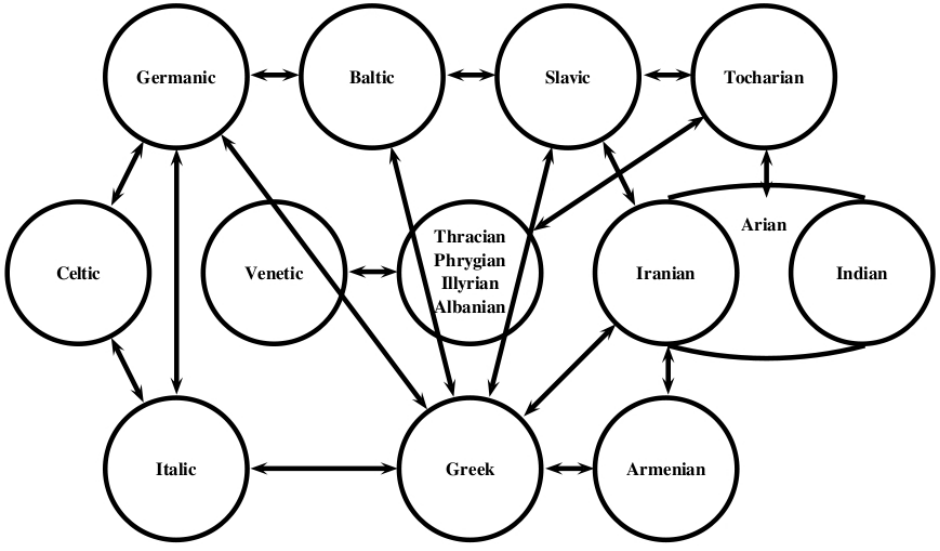
Meillet (1908)



Bloomfield (1933)



Hirt (1905)



Bonfante (1931)

Phylogenetic Networks

Trees are bad, because

Phylogenetic Networks

Trees are bad, because

- they are so difficult to reconstruct

Phylogenetic Networks

Trees are bad, because

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

Phylogenetic Networks

Trees are bad, because

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

Phylogenetic Networks

Trees are bad, because

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

Waves are bad, because

- nobody knows how to reconstruct them

Phylogenetic Networks

Trees are bad, because

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

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

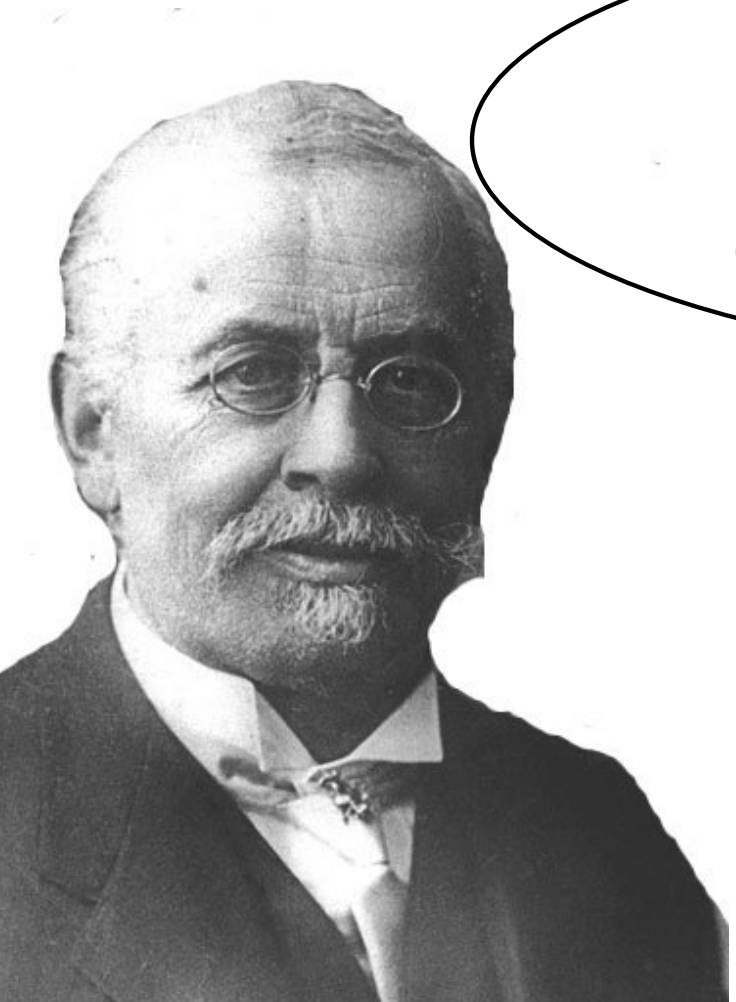
Waves are bad, because

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

Phylogenetic Networks

Hugo Schuchardt
(1842-1927)

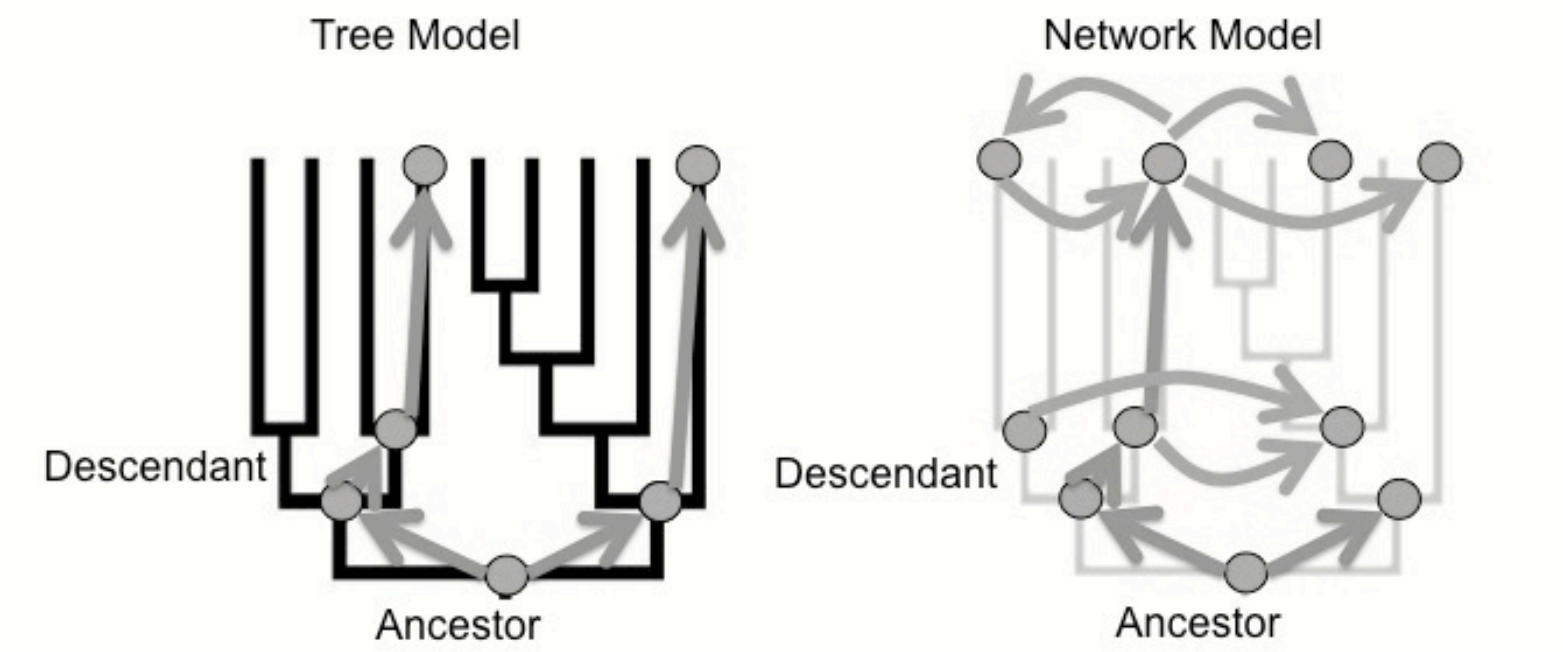
Phylogenetic Networks



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

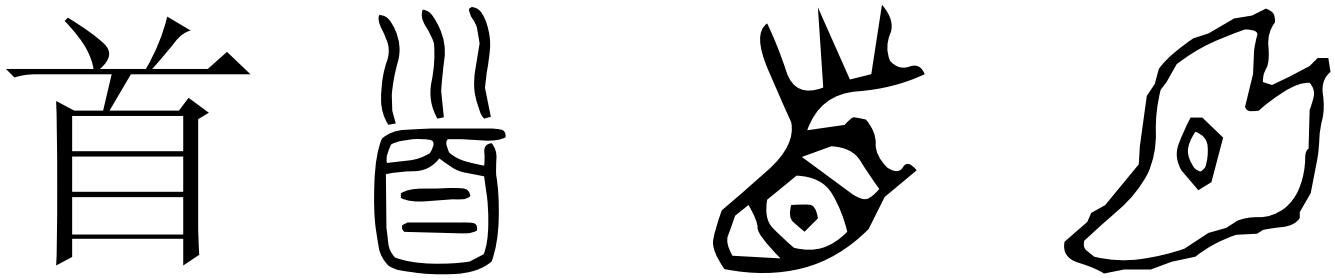
Hugo Schuchardt
(1842-1927)

Phylogenetic Networks



?

Modelling Chinese Dialect History



Data

Data

- Data was taken from the 现代汉语方言音库 Xiàndài Hànyǔ Fāngyán Yīnkù (Hóu 2004).

Data

- Data was taken from the 现代汉语方言音库 Xiàndài Hànyǔ Fāngyán Yīnkù (Hóu 2004).
- 180 *items* (“concepts”), translated into 40 dialect varieties of Chinese.

Data

- Data was taken from the 现代汉语方言音库 Xiàndài Hànyǔ Fāngyán Yīnkù (Hóu 2004).
- 180 *items* (“concepts”), translated into 40 dialect varieties of Chinese.
- Original source provides the data in RTF format (phonetic transcription, proposed underlying characters) along with audio files.

Data

- Data was taken from the 现代汉语方言音库 Xiàndài Hànyǔ Fāngyán Yīnkù (Hóu 2004).
- 180 *items* (“concepts”), translated into 40 dialect varieties of Chinese.
- Original source provides the data in RTF format (phonetic transcription, proposed underlying characters) along with audio files.
- RTF data was converted to text-format in order to allow automatic comparison.

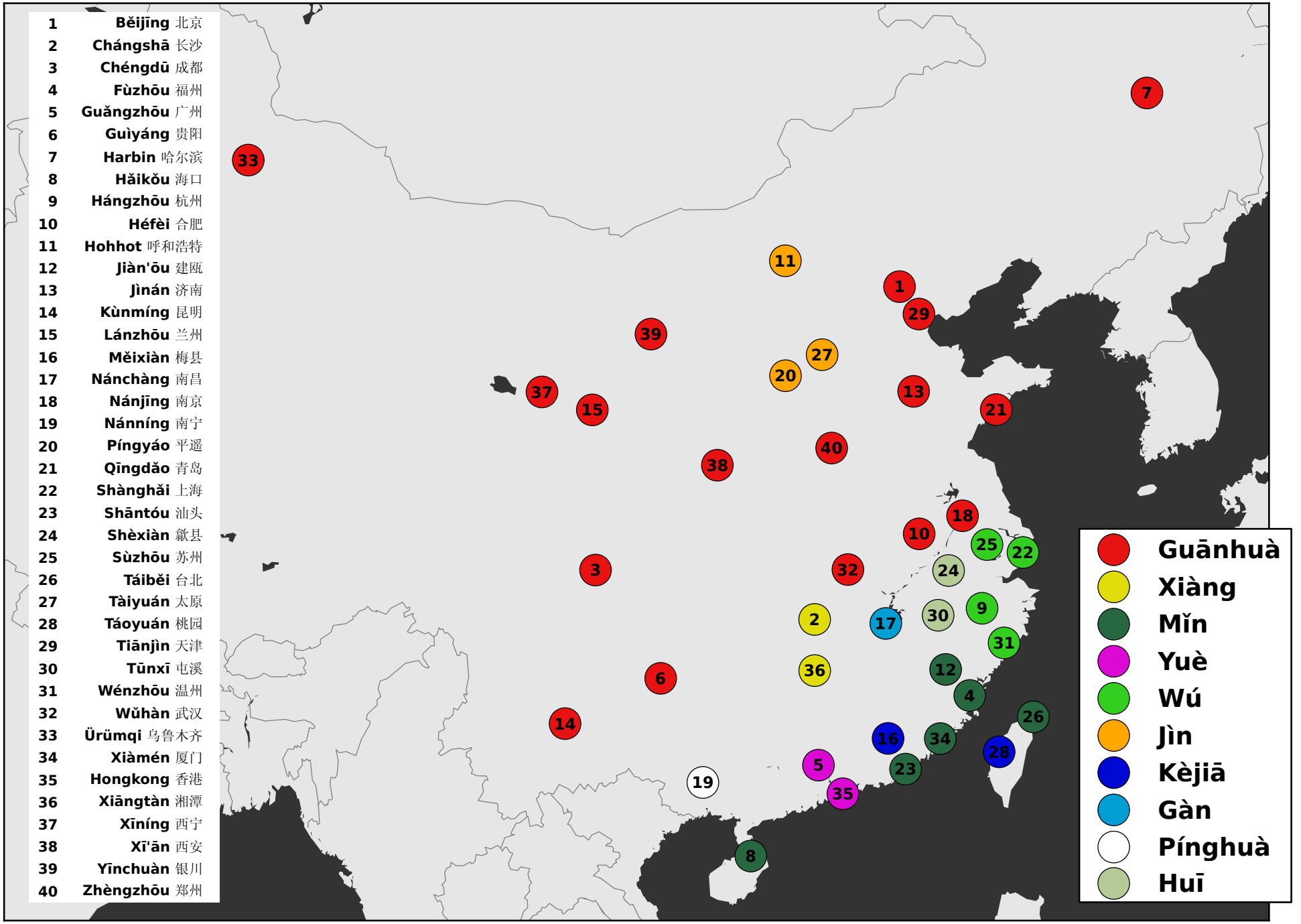
Data

- Data was taken from the 现代汉语方言音库 Xiàndài Hànyǔ Fāngyán Yīnkù (Hóu 2004).
- 180 *items* (“concepts”), translated into 40 dialect varieties of Chinese.
- Original source provides the data in RTF format (phonetic transcription, proposed underlying characters) along with audio files.
- RTF data was converted to text-format in order to allow automatic comparison.
- All entries were compared with the original transcriptions and the audio-files in order to decrease the number of errors that might have resulted from the conversion or the transcriptions.

Data

ITEM 太阳 *tàiyáng* “sun”

Dialect	Pronunciation	Character	Cognacy
上海 Shanghai	$t^ha^{34\text{-}33}ɦiã^{13\text{-}44}$	太阳	1
上海 Shànghǎi	$ȵjɪʔ^{1\text{-}11}dɤ^{13\text{-}23}$	日头	2
温州 Wénzhōu	$t^ha^{42\text{-}22}ji$	太阳	1
温州 Wénzhōu	$ȵi^{213\text{-}22}dɤu$	日头	2
广州 Guǎngzhōu	$jit^2t^hɐu^{21\text{-}35}$	热头	3
广州 Guǎngzhōu	$t^hai^{33}jœŋ^{21}$	太阳	1
海口 Hǎikǒu	zit^3hau^{31}	日头	2
北京 Běijīng	$t^hai^{51}iɑŋ^1$	太阳	1


1 Běijīng 北京
2 Chángshā 长沙
3 Chéngdū 成都
4 Fùzhōu 福州
5 Guǎngzhōu 广州
6 Guìyáng 贵阳
7 Harbin 哈尔滨
8 Hǎikǒu 海口
9 Hángzhōu 杭州
10 Héfèi 合肥
11 Hohhot 呼和浩特
12 Jiàn'ōu 建瓯
13 Jìnán 济南
14 Kùnmíng 昆明
15 Lánzhōu 兰州
16 Měixiàn 梅县
17 Nánchàng 南昌
18 Nánjīng 南京
19 Nánníng 南宁
20 Píngyáo 平遥
21 Qīngdǎo 青岛
22 Shànghǎi 上海
23 Shāntóu 汕头
24 Shèxiàn 歙县
25 Sùzhōu 苏州
26 Táiběi 台北
27 Tàiyuán 太原
28 Táoyuán 桃园
29 Tiānjìn 天津
30 Tūnxī 屯溪
31 Wénzhōu 温州
32 Wǔhàn 武汉
33 Ürümqi 乌鲁木齐
34 Xiàmén 厦门
35 Hongkong 香港
36 Xiāngtàn 湘潭
37 Xīníng 西宁
38 Xī'ān 西安
39 Yīnchuàn 银川
40 Zhèngzhōu 郑州
Guānhuà
Xiàng
Mǐn
Yuè
Wú
Jìn
Kèjiā
Gàn
Pínghuà
Huī

Analysis

Analysis

- The data was analysed with help of an improved version of the *minimal lateral network* approach (Dagan & Martin 2007, Dagan et al. 2008).

Analysis

- The data was analysed with help of an improved version of the *minimal lateral network* approach (Dagan & Martin 2007, Dagan et al. 2008).
- This version is freely available as part of a larger Python library for quantitative tasks in historical linguistics (LingPy, List & Moran 2013).

Analysis

- The data was analysed with help of an improved version of the *minimal lateral network* approach (Dagan & Martin 2007, Dagan et al. 2008).
- This version is freely available as part of a larger Python library for quantitative tasks in historical linguistics (LingPy, List & Moran 2013).
  - Starting from a reference tree that should display the “true” history of the languages as closely as possible, and a set of homologous characters (etymologically related words, cognates), the MLN approach infers horizontal relations between the contemporary and ancestral languages in the reference tree.

Analysis

- The data was analysed with help of an improved version of the *minimal lateral network* approach (Dagan & Martin 2007, Dagan et al. 2008).
- This version is freely available as part of a larger Python library for quantitative tasks in historical linguistics (LingPy, List & Moran 2013).
  - Starting from a reference tree that should display the “true” history of the languages as closely as possible, and a set of homologous characters (etymologically related words, cognates), the MLN approach infers horizontal relations between the contemporary and ancestral languages in the reference tree.
  - For each character (cognate set), a specific scenario which is closest to the patterns observed in the rest of the data is reconstructed.

Analysis

- The data was analysed with help of an improved version of the *minimal lateral network* approach (Dagan & Martin 2007, Dagan et al. 2008).
- This version is freely available as part of a larger Python library for quantitative tasks in historical linguistics (LingPy, List & Moran 2013).
  - Starting from a reference tree that should display the “true” history of the languages as closely as possible, and a set of homologous characters (etymologically related words, cognates), the MLN approach infers horizontal relations between the contemporary and ancestral languages in the reference tree.
  - For each character (cognate set), a specific scenario which is closest to the patterns observed in the rest of the data is reconstructed.
  - The main criterion for the selection of scenarios is homogeneity of the distribution of words across a fixed set of meanings in the sample.

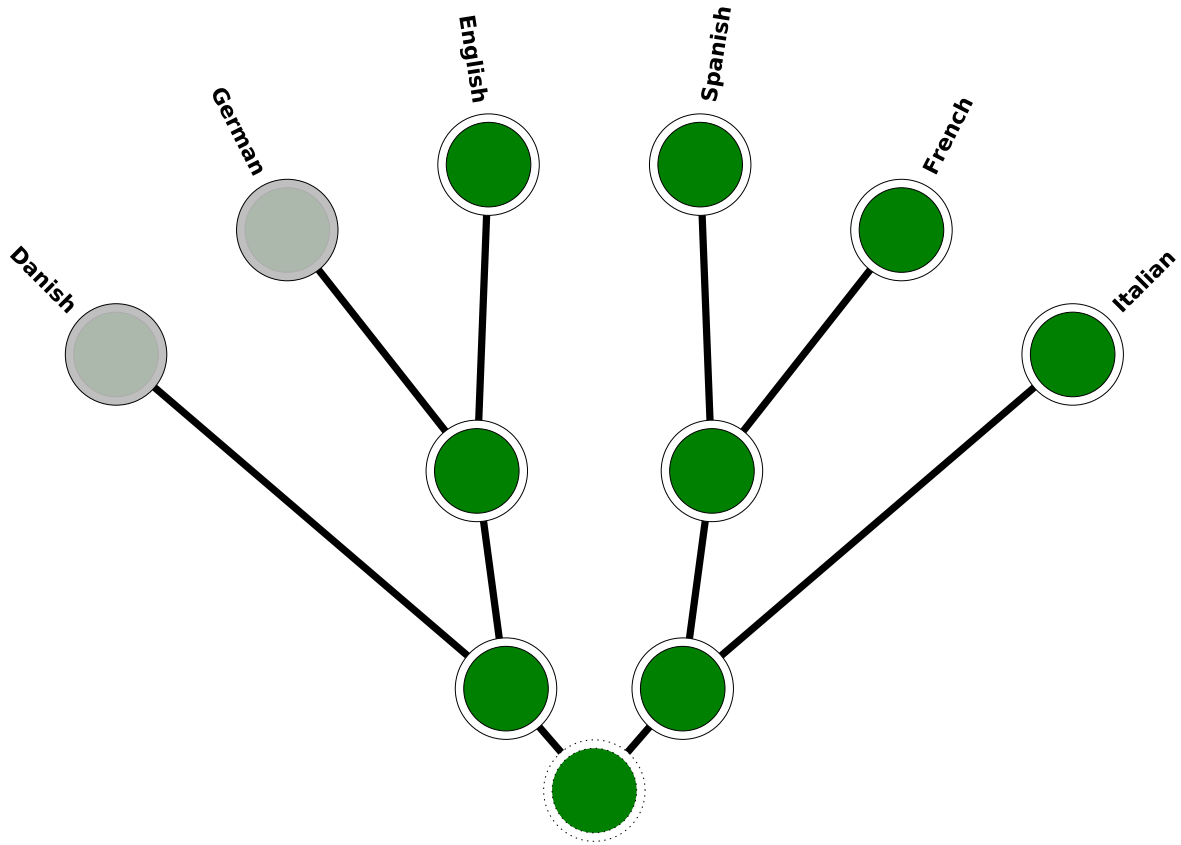
Analysis

- The data was analysed with help of an improved version of the *minimal lateral network* approach (Dagan & Martin 2007, Dagan et al. 2008).
- This version is freely available as part of a larger Python library for quantitative tasks in historical linguistics (LingPy, List & Moran 2013).
  - Starting from a reference tree that should display the “true” history of the languages as closely as possible, and a set of homologous characters (etymologically related words, cognates), the MLN approach infers horizontal relations between the contemporary and ancestral languages in the reference tree.
  - For each character (cognate set), a specific scenario which is closest to the patterns observed in the rest of the data is reconstructed.
  - The main criterion for the selection of scenarios is homogeneity of the distribution of words across a fixed set of meanings in the sample.
  - As a result, the method detects patterns that are *suggestive of borrowing* (patchy cognate sets). These can be directly reported to the researcher for further analysis or displayed in form of a rooted network.

Analysis

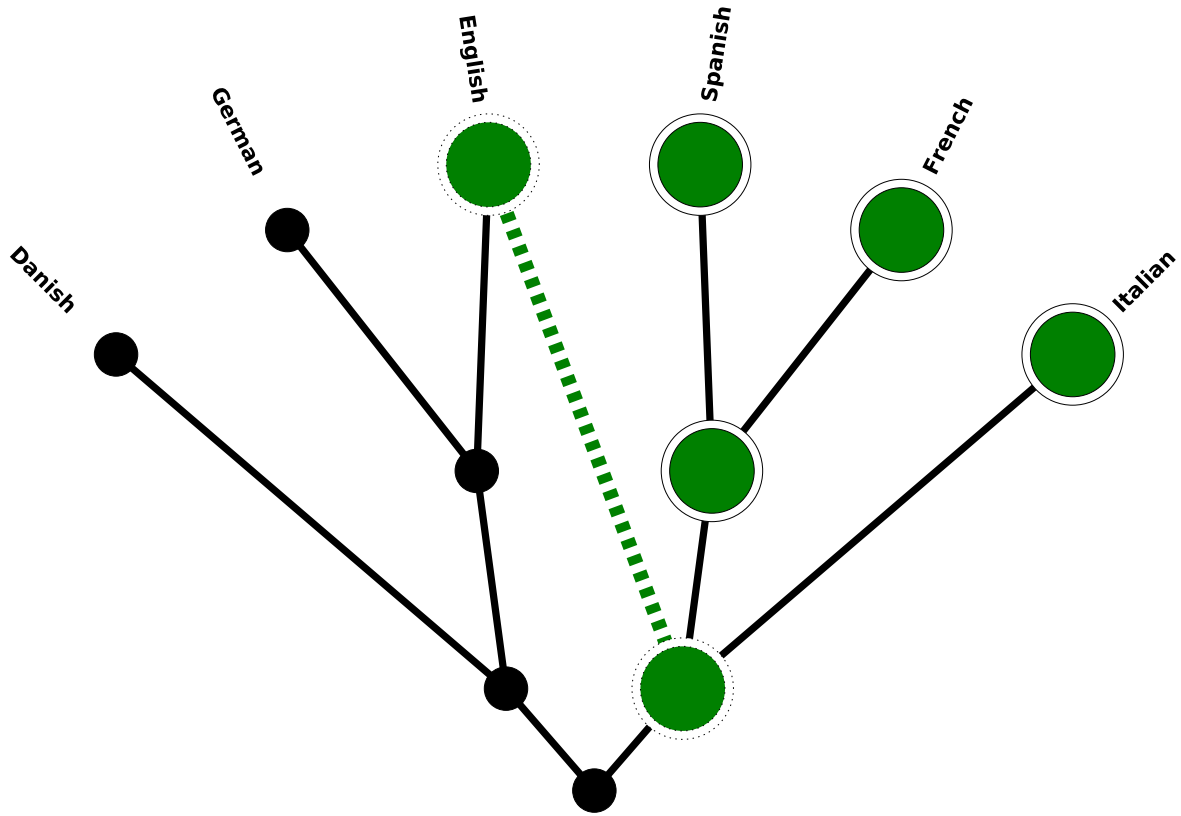
- The data was analysed with help of an improved version of the *minimal lateral network* approach (Dagan & Martin 2007, Dagan et al. 2008).
- This version is freely available as part of a larger Python library for quantitative tasks in historical linguistics (LingPy, List & Moran 2013).
  - Starting from a reference tree that should display the "true" history of the languages as closely as possible, and a set of homologous characters (etymologically related words, cognates), the MLN approach infers horizontal relations between the contemporary and ancestral languages in the reference tree.
  - For each character (cognate set), a specific scenario which is closest to the patterns observed in the rest of the data is reconstructed.
  - The main criterion for the selection of scenarios is homogeneity of the distribution of words across a fixed set of meanings in the sample.
  - As a result, the method detects patterns that are *suggestive of borrowing* (patchy cognate sets). These can be directly reported to the researcher for further analysis or displayed in form of a rooted network.
- The reference tree used for the analysis is based on Laurent Sagart's (pers. comm.) proposal for an innovation-based subgrouping of the Chinese dialects in which 瓦乡 Wǎxiāng and 蔡家 Càijiā (both not in our data) are taken as as primary branches.

Analysis



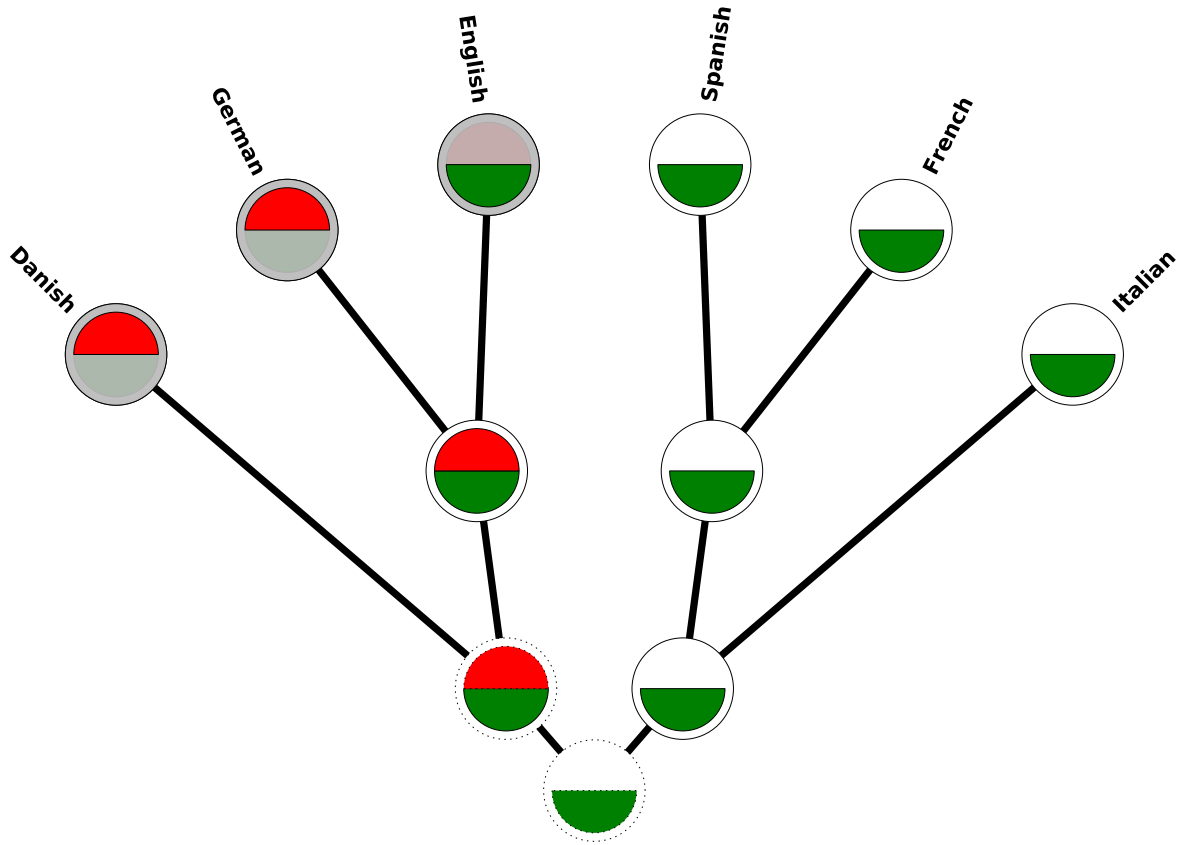
Language Variety	Danish	German	English	Spanish	French	Italian
"to count"	*tælle*	*zählen*	*count*	*contar*	*compter*	*contare*
Latin *computare*	0	0	1	1	1	1
Proto-Germanic **tal-*	1	1	0	0	0	0

Analysis



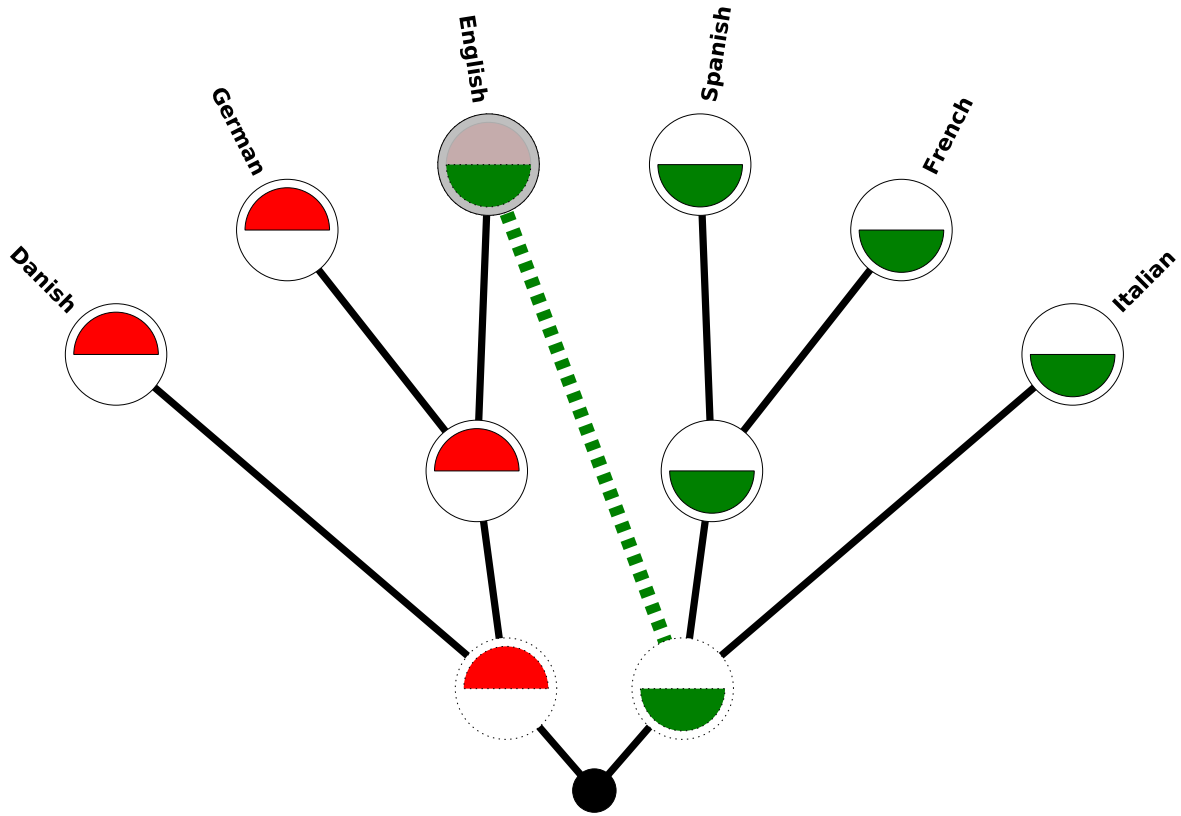
Language Variety	Danish	German	English	Spanish	French	Italian
"to count"	*tælle*	*zählen*	*count*	*contar*	*compter*	*contare*
Latin *computare*	0	0	1	1	1	1
Proto-Germanic **tal-*	1	1	0	0	0	0

Analysis



Language Variety	Danish	German	English	Spanish	French	Italian
"to count"	*tælle*	*zählen*	*count*	*contar*	*compter*	*contare*
Latin *computare*	0	0	1	1	1	1
Proto-Germanic **tal-*	1	1	0	0	0	0

Analysis



Language Variety	Danish	German	English	Spanish	French	Italian
"to count"	*tælle*	*zählen*	*count*	*contar*	*compter*	*contare*
Latin *computare*	0	0	1	1	1	1
Proto-Germanic **tal-*	1	1	0	0	0	0

Analysis

Sounds nice, but how *good* does the method work?

Analysis

Sounds nice, but how *good* does the method work?

- A test on 40 Indo-European languages showed that out of 105 cognate sets containing known borrowings, 76 were correctly identified as such.

Analysis

Sounds nice, but how *good* does the method work?

- A test on 40 Indo-European languages showed that out of 105 cognate sets containing known borrowings, 76 were correctly identified as such.
- Of 19 borrowings in English, 17 were correctly identified by the method.

Analysis

Ok, nice, but isn't there anything else you forgot to say?

Analysis

Ok, nice, but isn't there anything else you forgot to say?

- As our test on the Indo-European data revealed, the method does not only detect borrowings. It detects all kinds of errors in the data. Among these are:

Analysis

Ok, nice, but isn't there anything else you forgot to say?

- As our test on the Indo-European data revealed, the method does not only detect borrowings. It detects all kinds of errors in the data. Among these are:
  - Cases of parallel semantic shift that look like borrowings for the method.

Analysis

Ok, nice, but isn't there anything else you forgot to say?

- As our test on the Indo-European data revealed, the method does not only detect borrowings. It detects all kinds of errors in the data. Among these are:
  - Cases of parallel semantic shift that look like borrowings for the method.
  - Erroneous cognate judgments that also look like borrowings.

Analysis

Ok, nice, but isn't there anything else you forgot to say?

- As our test on the Indo-European data revealed, the method does not only detect borrowings. It detects all kinds of errors in the data. Among these are:
  - Cases of parallel semantic shift that look like borrowings for the method.
  - Erroneous cognate judgments that also look like borrowings.
  - Methodological errors (deep etymologies although the stochastic models require shallow ones, fuzzy concepts as basis, erroneous translations).

Analysis

Ok, nice, but isn't there anything else you forgot to say?

- As our test on the Indo-European data revealed, the method does not only detect borrowings. It detects all kinds of errors in the data. Among these are:
  - Cases of parallel semantic shift that look like borrowings for the method.
  - Erroneous cognate judgments that also look like borrowings.
  - Methodological errors (deep etymologies although the stochastic models require shallow ones, fuzzy concepts as basis, erroneous translations).
- It is certainly a benefit, that we can use the method to clean our data, but we should be careful with the results and only use it as an initial heuristic.

Results: General

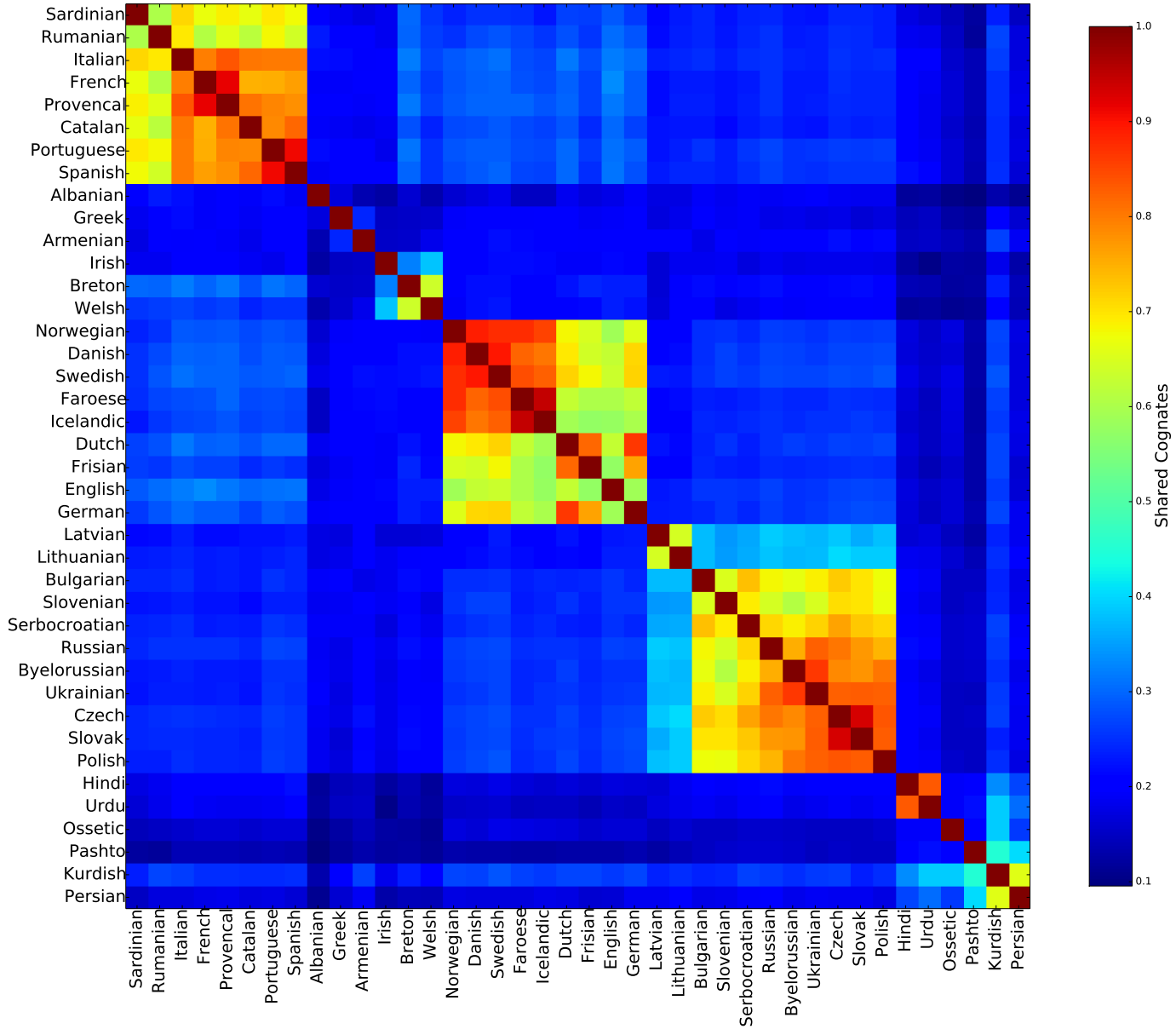
- 56% of the characters cannot be explained with help of the reference tree.

Results: General

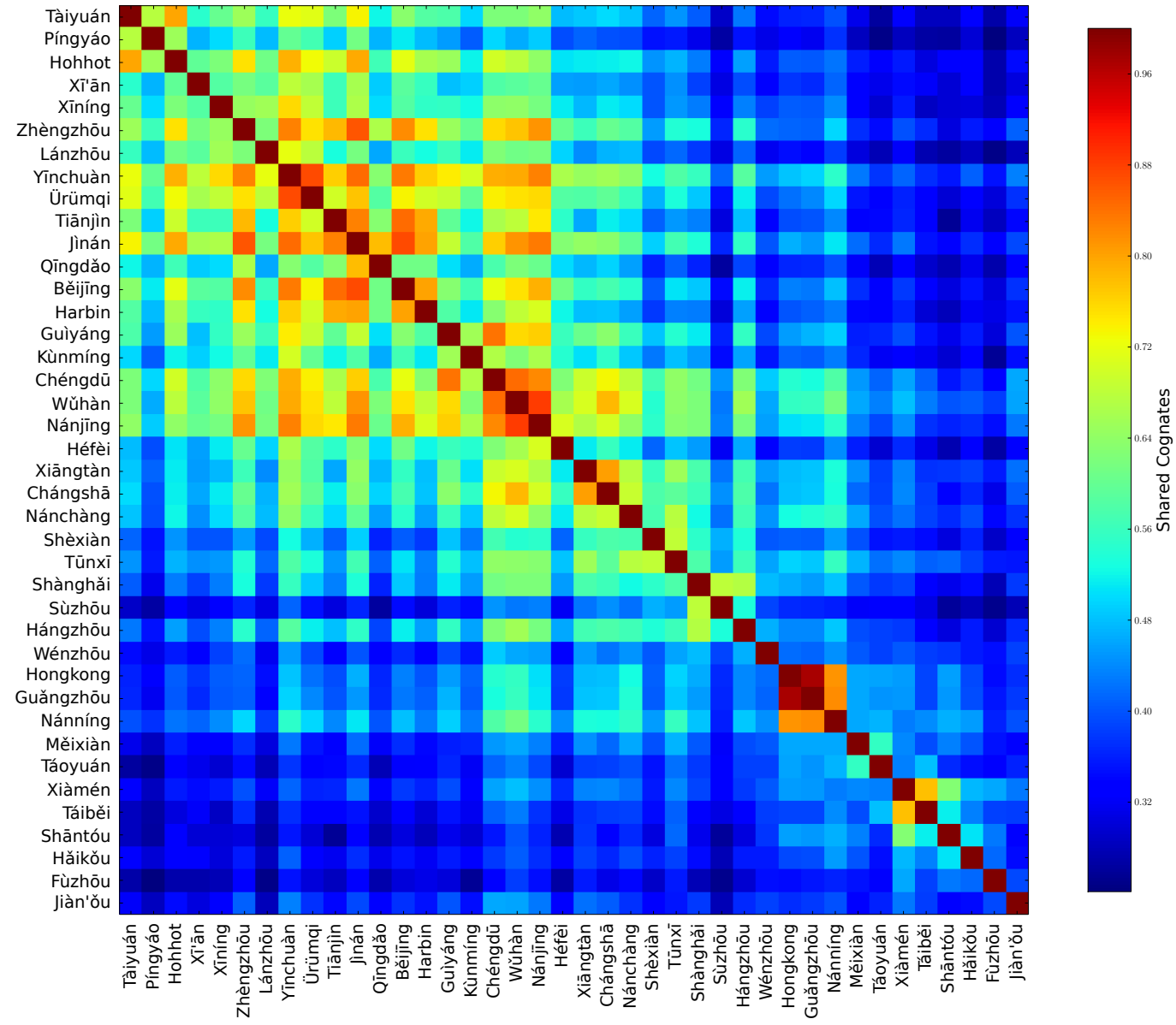
- 56% of the characters cannot be explained with help of the reference tree.
- This proportion is almost two times higher than was inferred for Indo-European (31%, 40 languages, 207 semantic items).

Results: General

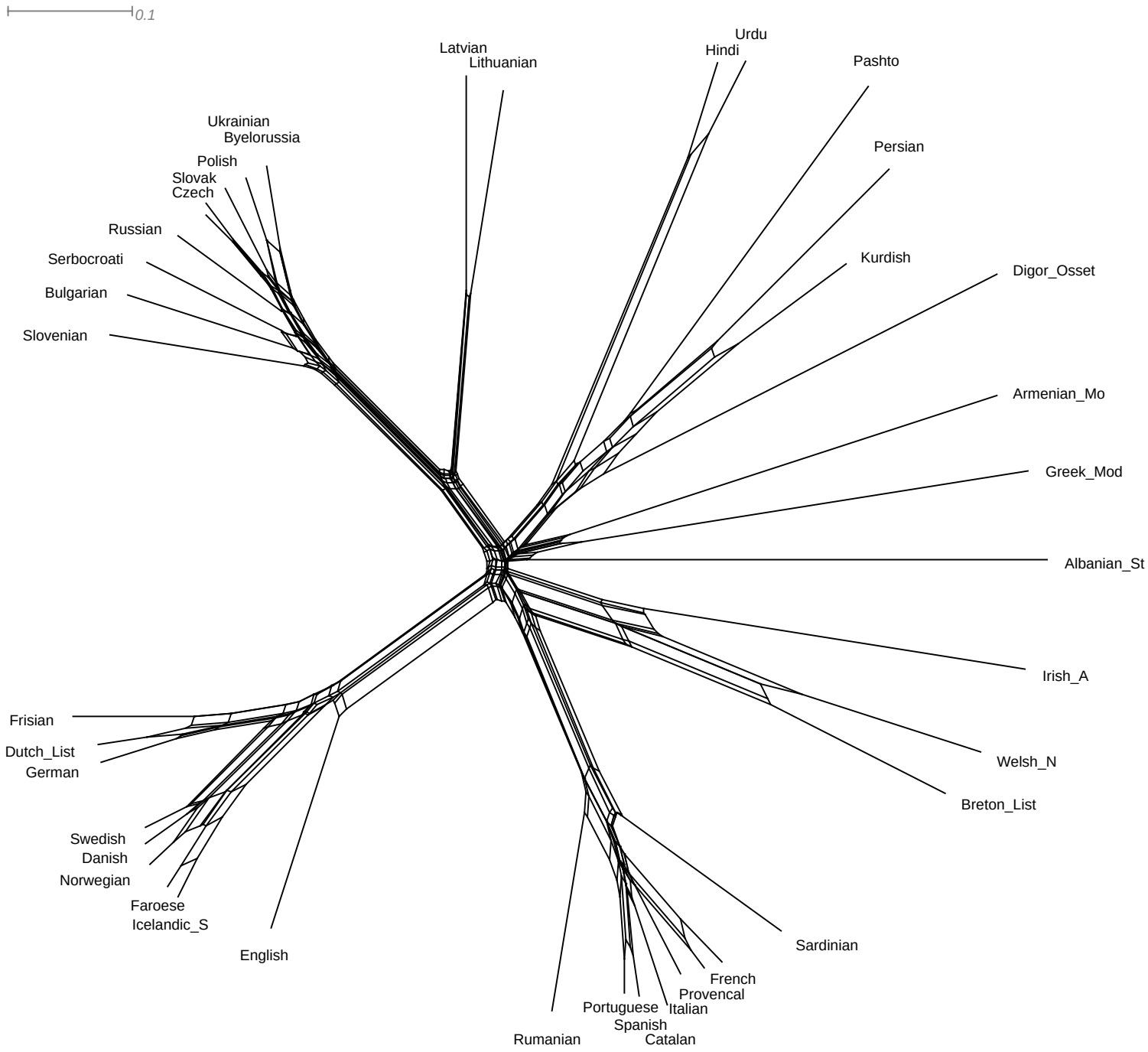
- 56% of the characters cannot be explained with help of the reference tree.
- This proportion is almost two times higher than was inferred for Indo-European (31%, 40 languages, 207 semantic items).
- Results might result from the fact that the concepts do not exclusively represent “basic concepts” (Swadesh 1952) and are thus more prone to borrowing. However, we don’t find a significant difference ($p = 0.16$, using Wilcoxon’s rank sum test) between between basic and non-basic concepts and the rest of the concepts.



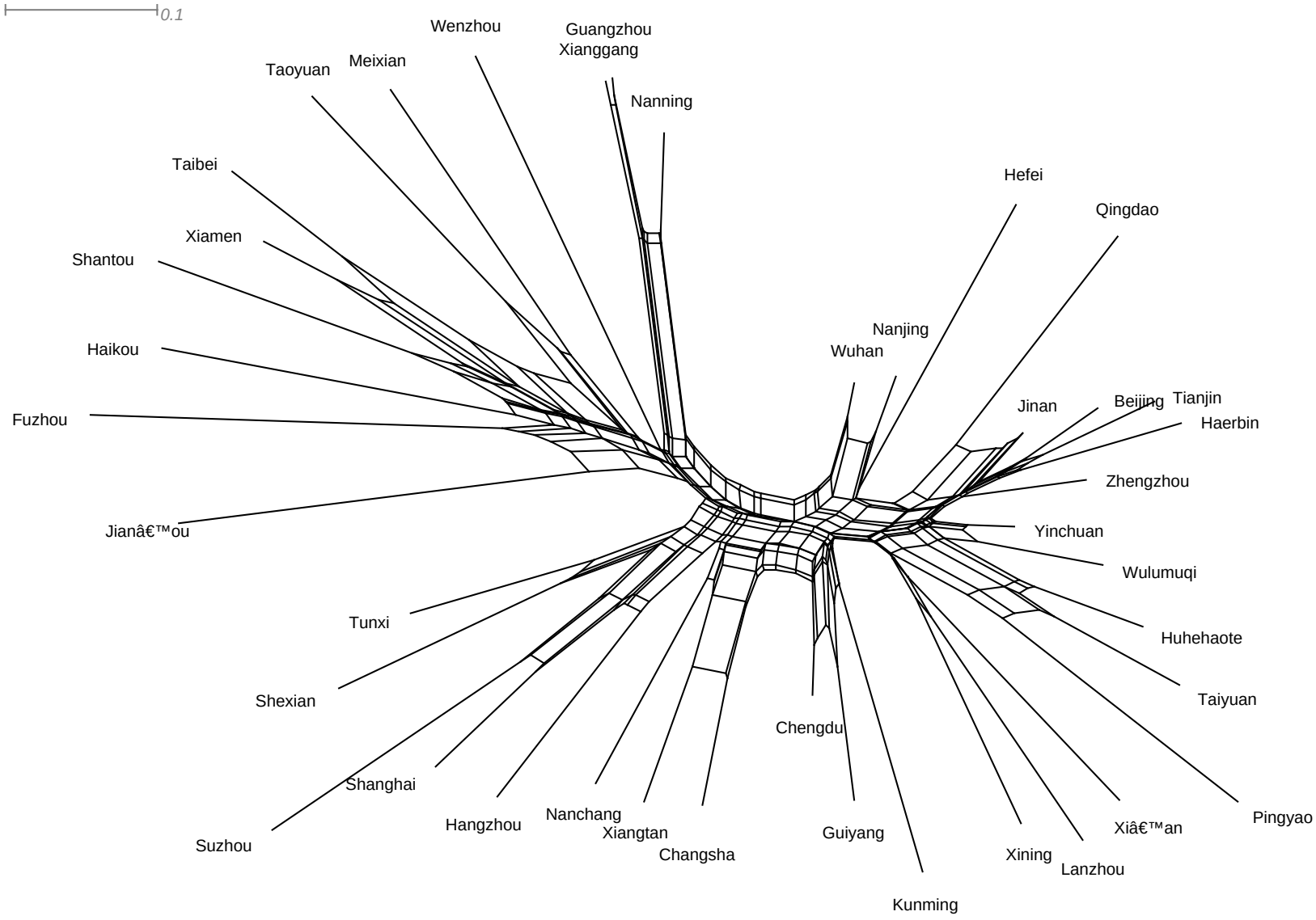
Shared cognate percentages (Indo-European)



Shared cognate percentages (Chinese)

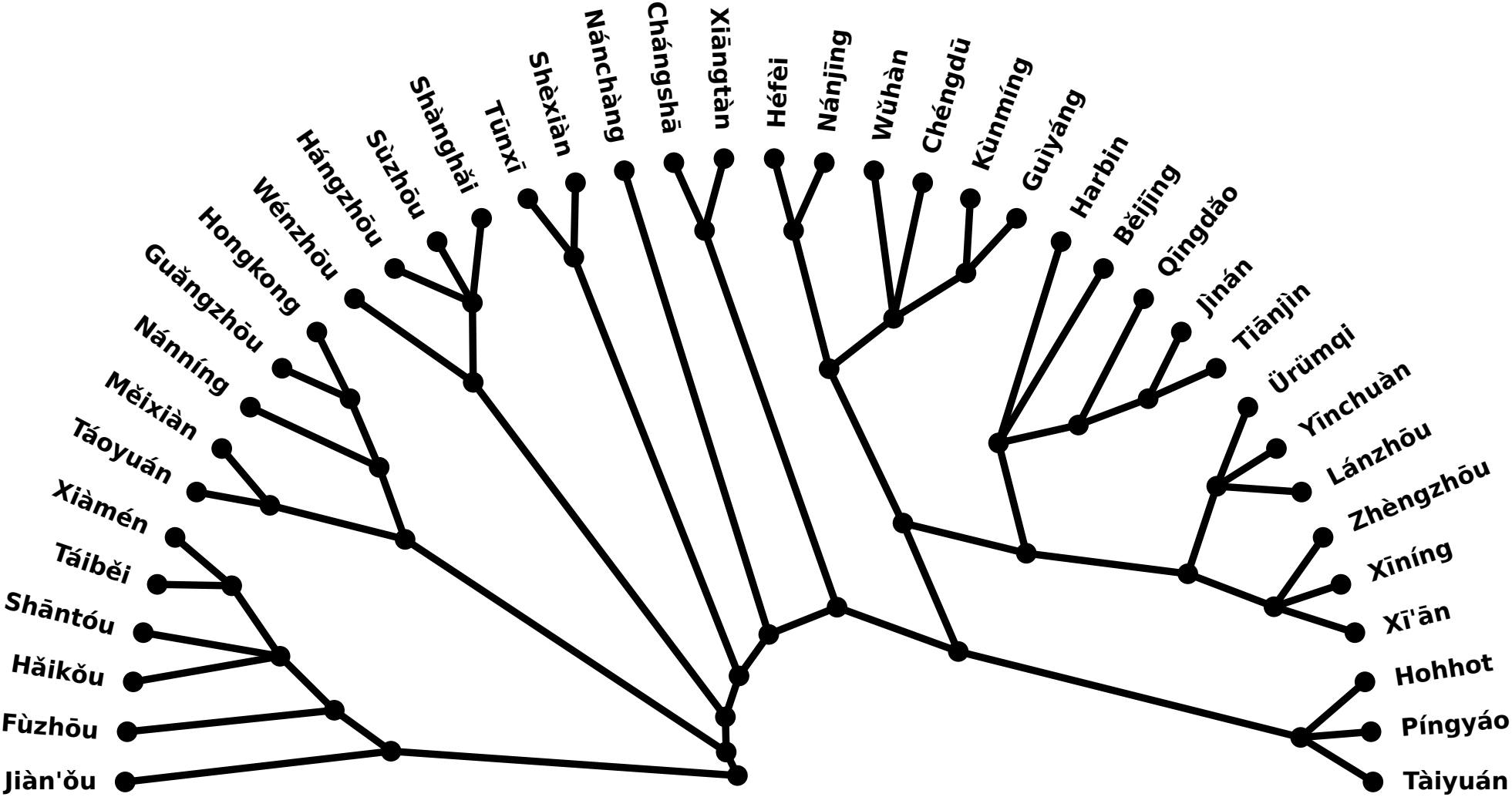


Neighbor-Net Analysis (Indo-European)



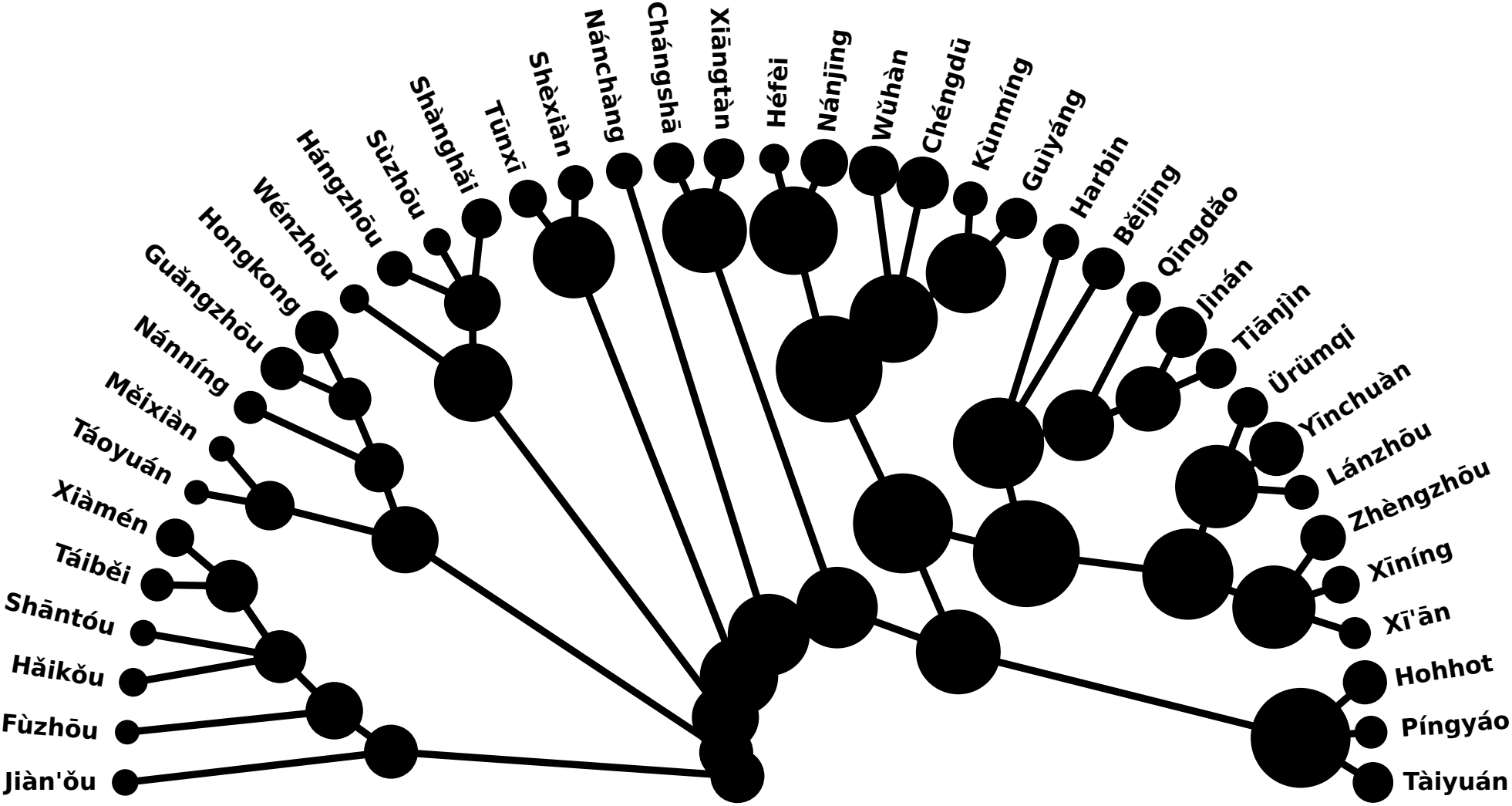
Neighbor-Net Analysis (Chinese)

Results: Minimal Lateral Network



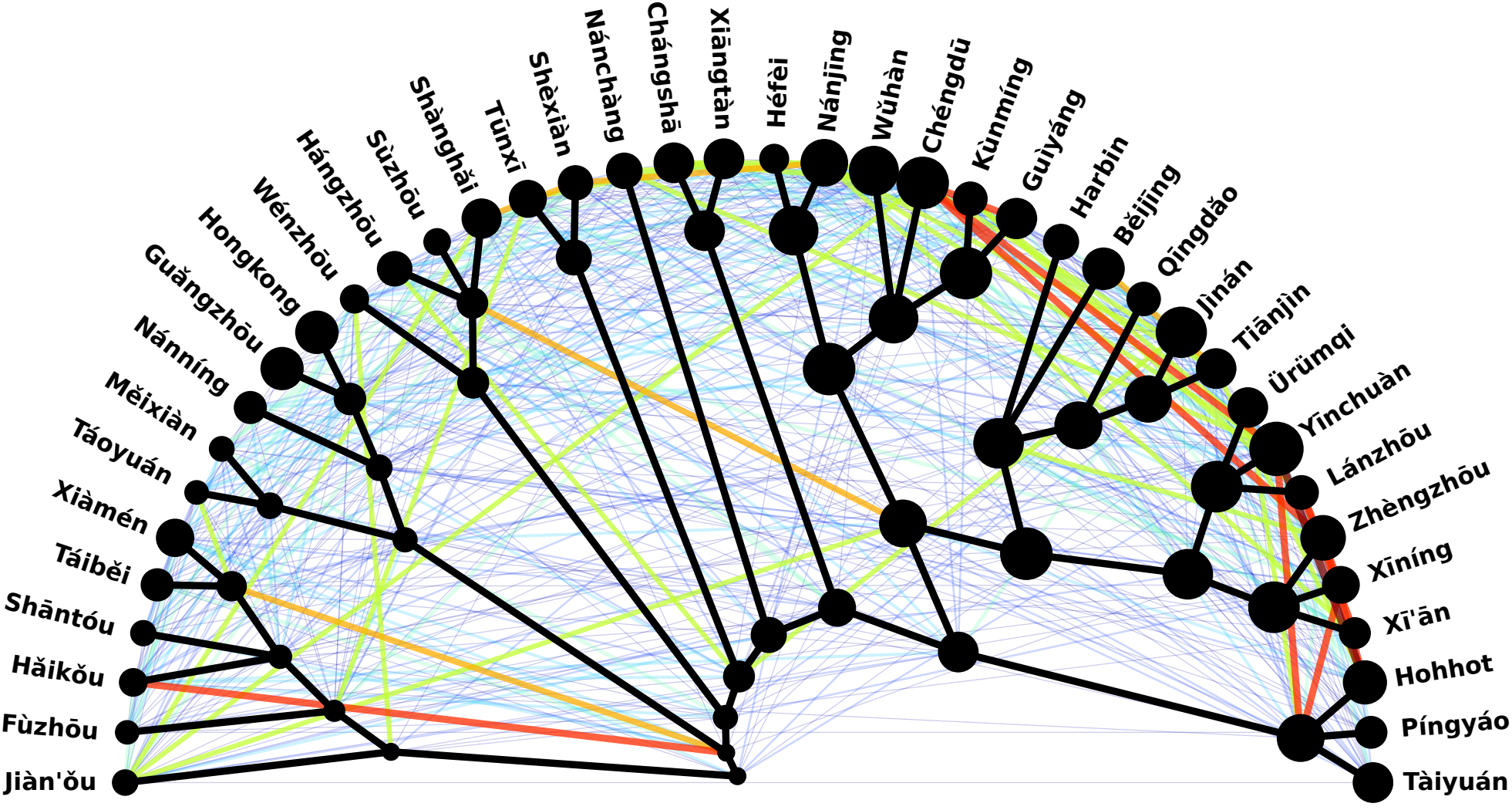
Reference tree of the Chinese dialects

Results: Minimal Lateral Network



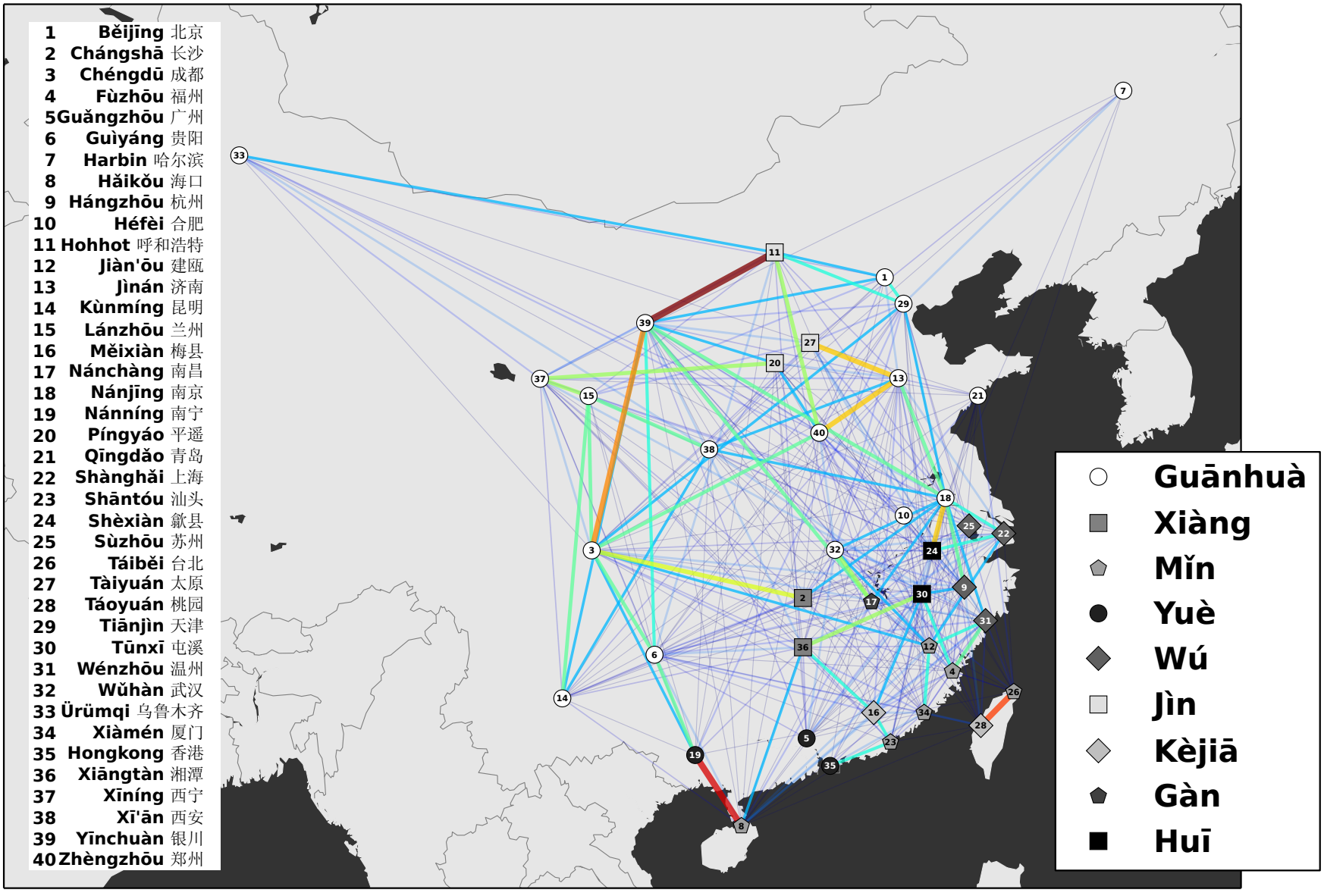
MLN analysis, no borrowing allowed

Results: Minimal Lateral Network

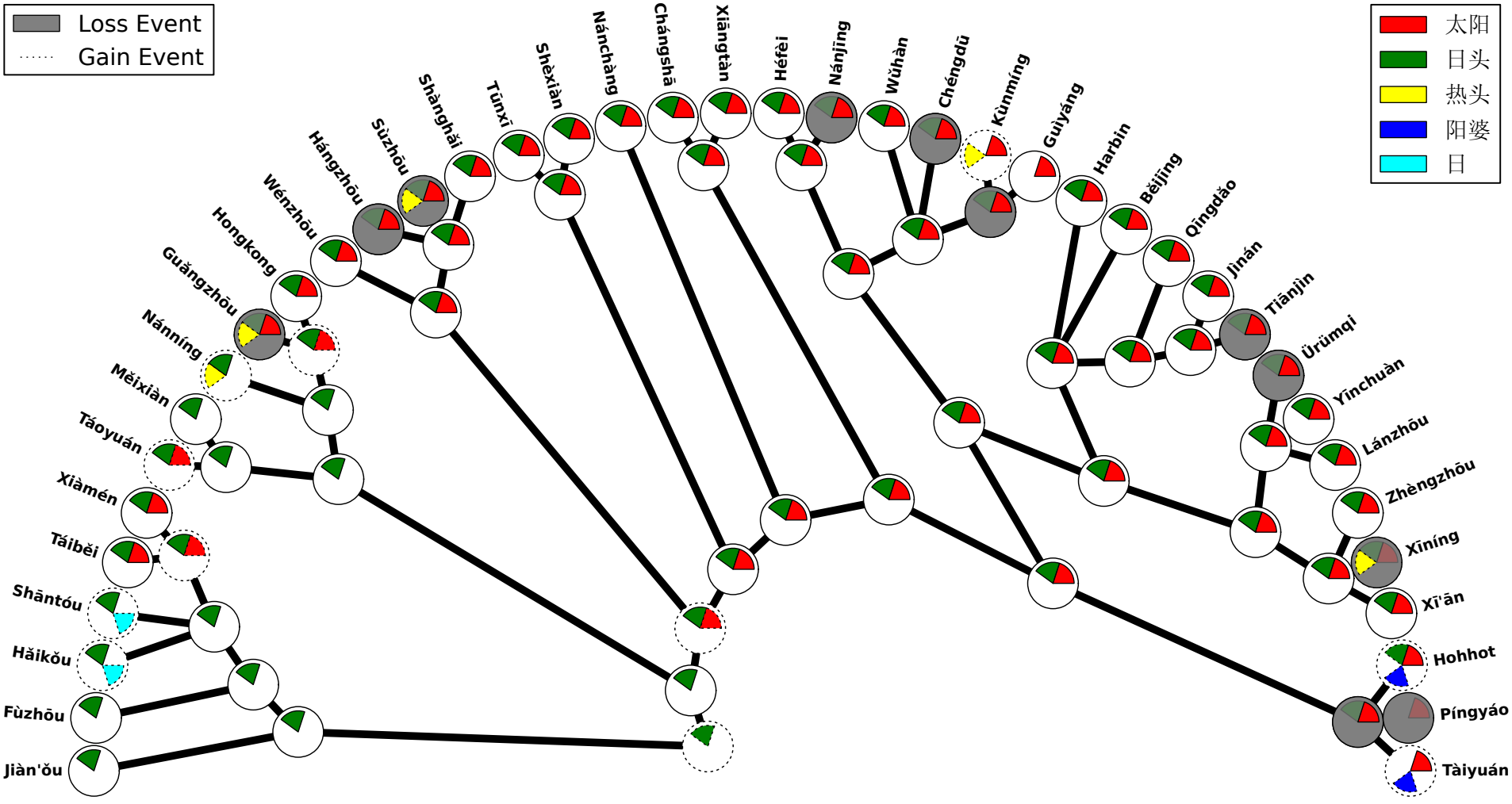


MLN analysis, best fit of borrowing and inheritance

Results: Minimal Lateral Network

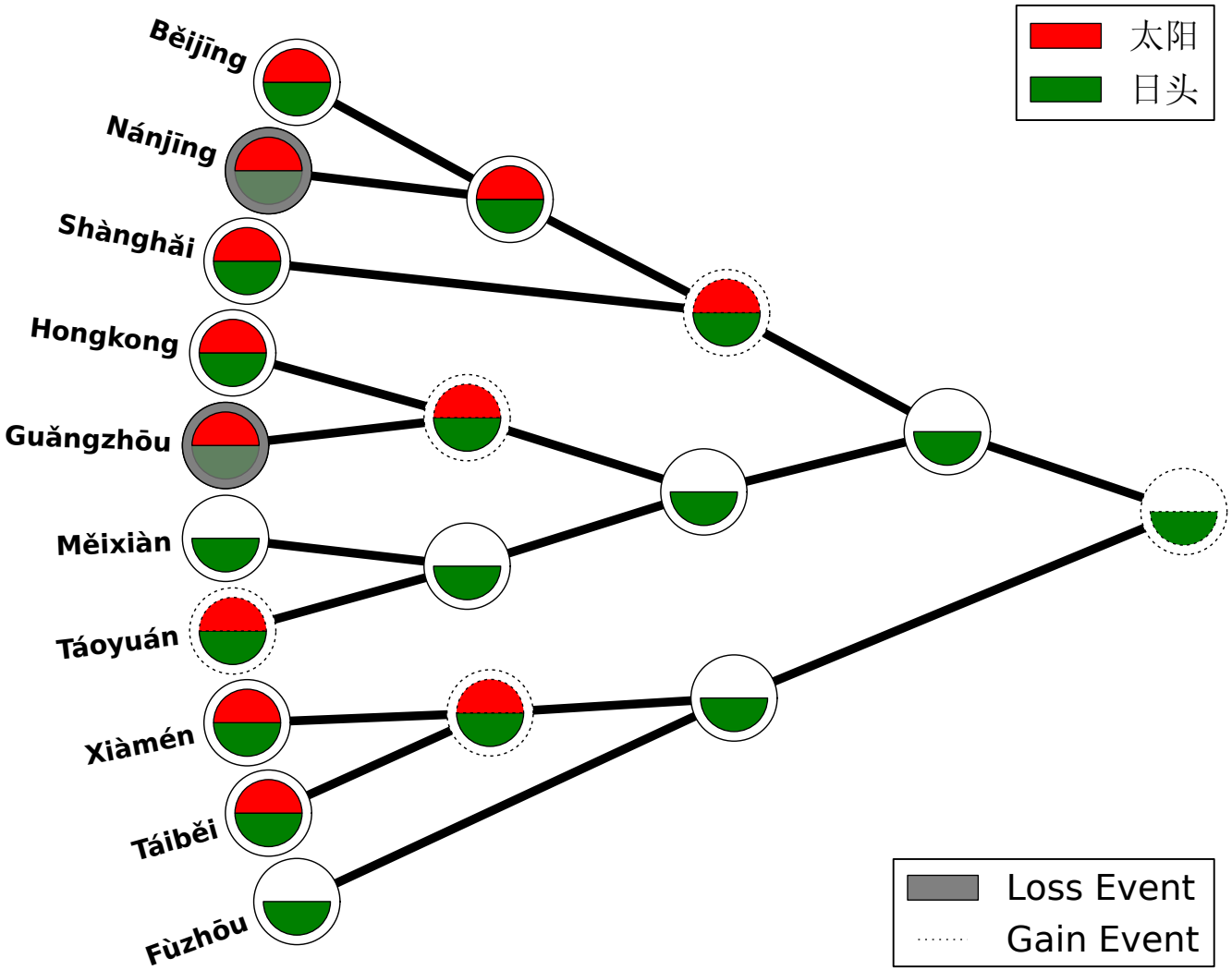


Results: Specific Scenarios



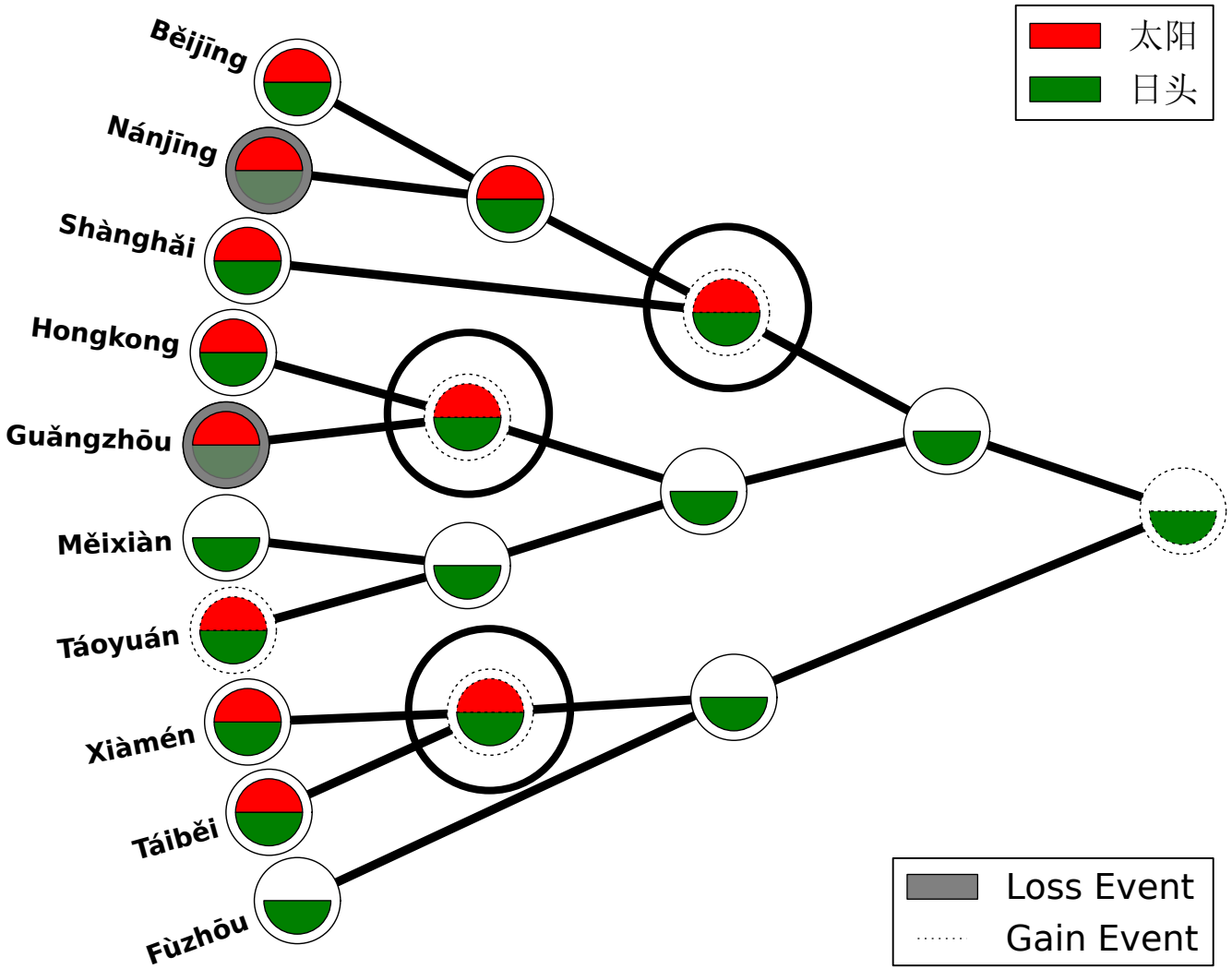
Item „sun”

Results: Specific Scenarios



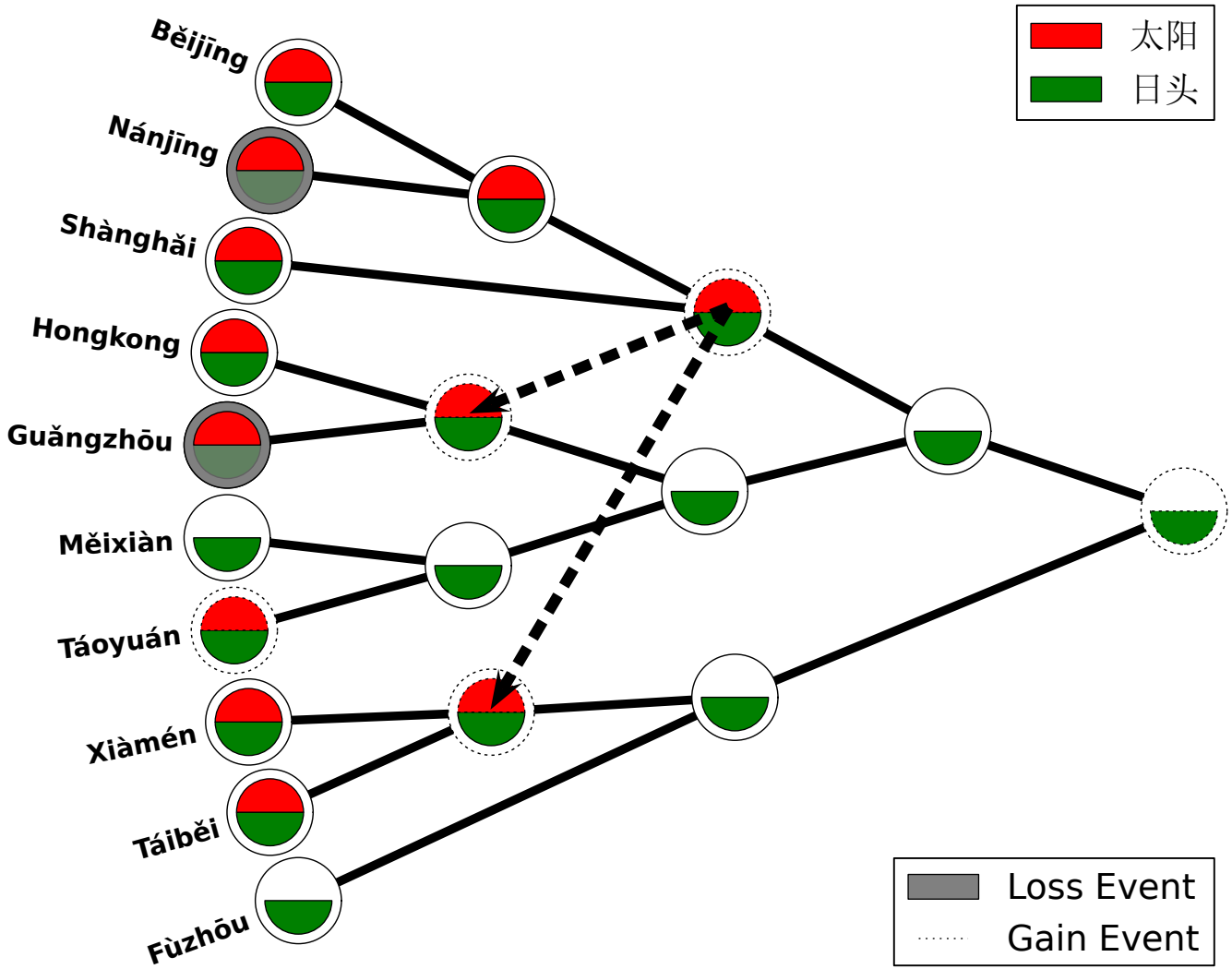
Item „sun”

Results: Specific Scenarios



Item „sun”

Results: Specific Scenarios



Item „sun”

Resultats: Specific Links

Node		Weight	Cognate Sets
Hǎikǒu	non-Mǐn	7	刚 刚 “just (just came)”, 淡 “light”, 南 瓜 “pumpkin”, 菠 菜 “spinach”, 勺 “spoon”, 瘦 “thin”, 从 “from”
Tàiběi, Xiàmén	non-Mǐn	6	只 “only”, 中 秋 节 “Mid-Autumn Festival”, 房间 “flat”, 只 classifier (cow), 冷 “cold”, 只 classifier (pig)
Tàiběi, Xiàmén	Táoyuán	6	豆 油 “soya sauce”, 包 仔 “baozi”, 太阳 “sun”, 桌仔“table”, 对 “from”, 看医生“go to the doctor”
Shànghǎi	Shèxiàn	6	彩虹 “rainbow”, 女人 “wife”, 爷 “father”, 落苏 “aubergine”, 山芋 “sweet potato”, 洋山芋 “spinach”
Hángzhōu	Mandarin, Huī, Xiàng, Gàn, Jìn	6	里头 “inside”, 哪个 “who”, 哪里 “where”, 那个 “that”, 刚好 “just right”, 包心菜 “cabbage”

Conclusion and Outlook

- Phylogenetic networks look nice.
- Phylogenetic networks can provide an alternative to both trees and waves.
- The application of phylogenetic network analyses in historical linguistics is still in its infancy. We have to test the methods further in order to get a better impression on its strong and weak points.

Conclusion and Outlook

谢谢大家！

Thank you!