BRILL

LANGUAGE DYNAMICS AND CHANGE 4 (2014) 222–252

brill.com/ldc

Using Phylogenetic Networks to Model Chinese Dialect History*

Johann-Mattis List
Forschungszentrum Deutscher Sprachatlas,
Philipps University Marburg, Marburg, Germany
mattis.list@uni-marburg.de

Shijulal Nelson-Sathi
Institute of Molecular Evolution,
Heinrich Heine University Düsseldorf, Düsseldorf, Germany
shijulalns@uni-duesseldorf.de

William Martin
Institute of Molecular Evolution,
Heinrich Heine University Düsseldorf, Düsseldorf, Germany
w.martin@uni-duesseldorf.de

Hans Geisler
Institute of Romance Languages and Literature,
Heinrich Heine University Düsseldorf, Düsseldorf, Germany
geisler@uni-duesseldorf.de

Abstract

The idea that language history is best visualized by a branching tree has been controversially discussed in the linguistic world and many alternative theories have been proposed. The reluctance of many scholars to accept the tree as the natural metaphor for language history was due to conflicting signals in linguistic data: many resemblances would simply not point to a unique tree. Despite these observations, the majority of automatic approaches applied to language data has been based on the tree model,

* First published in Søren Wichmann and Jeff Good (eds.). 2014. *Quantifying Language Dynamics: On the Cutting Edge of Areal and Phylogenetic Linguistics*, 125–154. Leiden: Brill.

 | DOI: 10.1163/22105832-00402008

while network approaches have rarely been applied. Due to the specific sociolinguistic situation in China, where very divergent varieties have been developing under the roof of a common culture and writing system, the history of the Chinese dialects is complex and intertwined. They are therefore a good test case for methods which no longer take the family tree as their primary model. Here we use a network approach to study the lexical history of 40 Chinese dialects. In contrast to previous approaches, our method is character-based and captures both vertical and horizontal aspects of language history. According to our results, the majority of characters in our data (about 54%) cannot be readily explained with the help of a given tree model. The borrowing events inferred by our method do not only reflect general uncertainties of Chinese dialect classification, they also reveal the strong influence of the standard language on Chinese dialect history.

Keywords

Chinese languages – Chinese linguistics – tree model – phylogenetic networks – lexical borrowing

1 Introduction

1.1 *Languages and Dialects*

What exactly is a language, and what is a dialect? One tends to say that the people from Shànghǎi, Běijīng, and Měixiàn all speak 'Chinese,' while people from Scandinavia speak 'Norwegian,' 'Swedish,' or 'Danish.' Looking at the phonetic transcriptions of the first sentence of Aesop's fable 'The Northwind and the Sun' in the three Chinese 'dialects' and the three Scandinavian 'languages' given in Table 1, the clear-cut distinction suggested by the different ways we name the varieties starts to become blurred. As the transcriptions show, the Chinese varieties differ from each other to a similar or even greater degree than the Scandinavian ones.

The reason for the fuzziness of the terms 'dialect' and 'language' can be found in the daily use of the terms in non-linguistic contexts. What is called a language and what a dialect does not necessarily depend on pure linguistic criteria, but often also on culture and politics (Barbour and Stevenson, 1998: 8). The problem of culture and politics, however, is that they have an impact on both languages and dialects. Although it certainly makes sense to state that Chinese dialects differ as much as the Scandinavian languages, it does not tell the whole truth about the sociolinguistic situation in China, where a large part

TABLE 1 *The first sentence of Aesop's fable 'The Northwind and the Sun' in different speech varieties. The words are semantically aligned, i.e. all translational equivalents are placed in the same column. Words shaded in gray are etymologically related*

Běijīng Chinese	iou^{21}	i^{55}	xuei^{35}	pei^{21}fəŋ^{55}	kən^{55}	tʰai^{51}iaŋ^{11}	ʈʂəŋ^{55}	tsai^{53}	
Měixiàn Chinese	iu^{33}	it^{55}	pai^{33}a^{11}	pet^{33}fuŋ^{33}	tʰuŋ^{11}	ɲit^{11}tʰeu^{11}	hɔk^{33}		
Shànghǎu Chinese	ɦi^{22}		tʰã^{55} tsɪ^{21}	poʔ^{3}foŋ^{44}	taʔ^{5}	tʰa^{33}ɦiã^{44}	tsəŋ^{33}	hɔ^{44}	

Běijīng Chinese (*cont.*)	naə^{51}	ʈʂəŋ^{55}luən^{51}		
Měixiàn Chinese	e^{53}	au^{55}		
Shànghǎu Chinese		ləʔ^{11}lə^{23}tsa^{53}		

Norwegian	nuːɾɑʊinˑɳ	ɔ	suːlɳ					kɾɑŋlət	ɔm
Swedish	nuːɖanvɪndən	ɔ	suːlən		tʏɻstadə	ən gɔŋ			ɔm
Danish	noˀʌnvenˀɳ	ʌ	soːlˀn	kʰʌm		eŋg̊ɑŋ	i sg̊ʁiðˀ		ʌmˀ

of the population is bilingual, using a common language for writing and—if necessary—also for verbal communication. In order to describe such complex heterogeneous structures as modern languages, sociolinguists have proposed the *diasystem* model (Branner, 2006: 209). According to this model, a language is a complex aggregate of different linguistic systems coexisting and mutually influencing each other (Coseriu, 1973: 40). Usually, a diasystem is determined by a *Dachsprache* (*roof language*), a linguistic variety that serves as a standard for interdialectal communication (Goossens, 1973: 11).

In the case of the Chinese diasystem, the *Dachsprache* is the modern standard language (henceforth called Standard Chinese), which was originally derived from the dialect of Běijīng, but—being used as second language throughout China—has long started to live a life of its own. Its influence can be noticed in almost all dialects. Lexically, it often appears in terms of multiple words for a single concept, with one representing the word originally used in the dialect, and one being borrowed from Standard Chinese. In the example given in Table 1, for example, Shànghǎi [tʰa^{33}ɦiã^{44}] 'sun' has been borrowed from Standard Chinese 太阳 *tàiyáng* [tʰai^{51}iaŋ^{11}] 'sun.' This can be seen from the fact that there is another word for 'sun' in Shànghǎi: [ɳjɪʔ^{11}dɤ^{23}]. This word is much older than the former and is cognate with Měixiàn [ɲit^{11}tʰeu^{11}] 'sun.' Cases where dialects borrow from the *Dachsprache* are very frequent in almost all Chinese dialects, while cases of borrowing between neighboring dialects are probably even more frequent.

1.2 *Trees, Waves, and Networks*

Ever since August Schleicher first proposed the idea that the evolution of languages is best visualized by a branching tree ('dem Bilde eines sich verästelnden Baumes'; Schleicher, 1853: 787), this view has been controversially discussed in the linguistic world, leading to various opposing theories ranging from wave-like evolutionary scenarios (Schmidt, 1872) to early network proposals (Bonfante, 1931). Since most alternative approaches remained static, disregarding the time dimension in favor of the spatial dimension, the tree was never completely abandoned, and both the family tree (*Stammbaum*) and the wave theory (*Wellentheorie*) became standard models of language change that were used interchangeably, depending on the respective questions that scholars wanted to elaborate. Although, during the history of linguistics, the idea of combining both models into a single framework was often discussed (Schuchardt, 1900; Southworth, 1964), linguists failed to propose a formal model for phylogenetic networks that would have allowed both vertical and horizontal language relations to be captured. As historical linguistics took a quantitative turn at the beginning of the third millennium, many methods that had originally been designed to model and infer biological evolution were repeatedly applied to linguistic problems. While most of these approaches continued with the tree model, comparing languages with species (Gray and Atkinson, 2003; McMahon and McMahon, 2005; Atkinson and Gray, 2006), recent research has shown (Nelson-Sathi et al., 2011, List et al., 2014) that network approaches originally used to model microbial evolution (Dagan and Martin, 2007; Dagan et al., 2008) might be even more apt for modeling language history. Network approaches not only offer a formal way to model vertical and horizontal language relations, but also provide different methods for inferring these relations from linguistic data. So far, however, phylogenetic network approaches are still in their infancy, both with respect to the methods that have been proposed and with respect to their applications.

The Chinese dialects seem to be a good test case for these new approaches. Given their complex history, their 'close proximity to one another for two millennia and the pervasive influence of various quasi-standards and koinés on all Chinese dialects over a very long period' (Norman, 2003: 76), it is obvious that they are 'not entirely amenable to a Stammbaum formulation' (ibid.). Here we apply a network approach to model the history of 40 Chinese dialect varieties. In contrast to previous network analyses of Chinese dialects that were based on split distances and only measured the uncertainty of trees (Ben Hamed and Wang, 2006), our approach is character-based: it automatically infers hidden borrowings in the data and thus captures both the vertical and horizontal aspects of language history.

2 Materials

2.1 *Data*

The data that we used for our analysis is taken from the *Hànyǔ Fāngyán Yīnkù* (Hóu, 2004), a CD-ROM that offers different resources for Chinese dialects including phonological descriptions, phonetic transcriptions, and sound recordings for 40 different dialect varieties. From the CD-ROM we extracted a lexical subset, consisting of 180 glosses ('concepts') translated into the respective varieties. Chinese dialects often have multiple synonyms for one concept; therefore the resulting dataset comprises 10,201 words. Since the word lists were compiled for dialect studies where the selection of lexical items is usually based on phonetic criteria, only 48 of the 180 glosses (26%) belong to the basic vocabulary in the strict sense of Swadesh (1952 and 1955). The source material was obtained in a format not suitable for computational analyses, requiring the extraction procedure to be carried out semi-automatically, with additional manual cleaning by the researchers/present authors. All entries were double-checked by comparing the phonetic transcription for each word with its corresponding sound recording. The data was further enriched by looking up the geographic coordinates of the central cities where the varieties are spoken, translating the glosses into English, adapting the phonetic transcriptions to standard IPA, and applying a rough procedure for automatic cognate detection that is described in detail in the following section. Table 2 shows an excerpt of the data in its current format.

TABLE 2 *The basic format of the input data*[1]

ID	Variety	Concept	St. Chinese	IPA	Char.	Cogn. Set
1	Shànghǎi	'sun'	*tàiyáng* 太阳	$t^ha^{34-33}ɦiã^{13-44}$	太阳	2
2	Shànghǎi	'sun'	*tàiyáng* 太阳	$ȵjɪʔ^{1-11}dɤ^{13-23}$	日头	1
3	Sūzhou	'sun'	*tàiyáng* 太阳	$ȵiəʔ^{3}dʏ^{13-21}$	热头	3
4	Sūzhou	'sun'	*tàiyáng* 太阳	$t^hɑ^{513-55}ɦiã^{13-21}$	太阳	2
5	Hángzhōu	'sun'	*tàiyáng* 太阳	$t^hɛ^{445}ɦiɑŋ^{213-31}$	太阳	2
6	Wēnzhōu	'sun'	*tàiyáng* 太阳	t^ha^{42-22}ji	太阳	2

2.2 *Cognate Judgments*

Along with the recent quantitative turn in historical linguistics, one can also observe a shift from the interest in *proto-forms* to an interest in *cognates*. This likewise holds for our approach, which requires sets of cognate words as input data. Cognates are usually defined as words or morphemes that are derived from a common ancestor form via vertical inheritance (Trask, 2000: 62). Our input requirements are less strict, however: the method only requires that the words are etymologically related, or *homolog* in the biological sense, i.e. that they share a common ancestry, no matter whether this is due to vertical transfer or borrowing (Koonin, 2005: 311). In Chinese dialectology, it is common to specify not only the pronunciation of a given dialect word, but also give an assessment regarding its homology. Homology assessments are usually coded by providing the Chinese characters corresponding to a given word.[2] Since for most Chinese characters the Middle Chinese readings (spoken around the 6th century CE) can be reconstructed from old rime books, a character is somewhat similar to a proto-form. Thus, Táoyuán [ŋit^{22}tʰeu^{11}] and Hǎikǒu [zit^{3}hau^{31}] 'sun' are both written as 日头, and the proto-form would have been pronounced as *ɳit^{4}duw^{1} in Middle Chinese times (if the compound was already present during that time).[3] Note that the character assignments in Chinese dialectology are homologs in the strict sense, since no distinction is drawn between borrowing and vertical inheritance.

While the postulation of a proto-form for a given set of words is—ideally—a full statement regarding their phonetic and phylogenetic history, being a shortcut formulation for known, regular sound change processes, the postulation of cognate relations between words is much simpler, being merely a statement that there *is* a history relating them. It is usually emphasized that the nature of this history should only involve vertical transmission. The details of vertical transmission are usually ignored, and no further distinction between the

1 Note that the character assignment correctly claims that Sūzhou [ɳiəʔ^{3}dʏ$^{13-21}$] and Shànghǎi [ɳjɪʔ$^{1-11}$dɤ$^{13-23}$] are not cognate, with the initial syllable of the former going back to Middle Chinese *ɳɐt 'hot' and the initial syllable of the latter going back to Middle Chinese *ɳit 'sun'. The words are, however, closely related, since it is not impossible that the original form in Sūzhou was a reflex of Middle Chinese *ɳit 'sun', but was later reinterpreted as Middle Chinese *ɳɐt 'hot'. However, this does not influence our strict criterion for cognacy assignments.

2 The procedure for choosing the characters is not always clear-cut. See Kurpaska (2010: 118–120) for details.

3 Middle Chinese character readings follow an IPA adaptation of the system of Baxter (1992).

different types is drawn. Thus, in lexicostatistical databases, such as the *Tower of Babel Database* (http://starling.rinet.ru) or the *Indo-European Lexical Cognacy Database* (http://ielex.mpi.nl/), the Italian and French words for 'give,' *dare* and *donner* respectively, are usually placed in the same cognate set, although they go back to two different Latin words (*dare* 'give' and *dōnare* 'give as a present'). The reason for this cognate assignment is that the Latin forms themselves go back to a common Indo-European root, with *dare* being a reflex of Proto-Indo-European *deh$_3$- 'give' and *dōnare* being a reflex of its nominalized form *deh$_3$-no- 'what is given' (cf. Meiser, 1998). Trask (2000: 234 f.) proposes the term *oblique cognates* to address these specific cases of indirect cognate relations, but the term is rarely used in the literature, and direct and indirect cognacy are usually treated identically in practice.

Another problem of cognate assignment that is ignored in most quantitative approaches is the problem of *partial cognacy*. Is it justified to say that compound words such as Spanish *porque* and Russian *potomu čto* 'because' are cognate, since certain parts of them (*-que* and *čto*) can be traced back to Proto-Indo-European **kwi-* 'what'? And, if so, what is their relation when adding more words to the comparison, such as Danish *fordi* 'because,' which is partially cognate with the Spanish word (*for-* ≈ *por-*) but not the Russian? In most datasets, this problem is solved by assigning compound words to multiple cognate sets, one for each morpheme. Such an approach, however, can become problematic when dealing with languages where compounding is frequent. In Table 3, the words denoting 'moon' in seven Chinese dialects are contrasted in such a way that all cognate morphemes are aligned, with the characters in the first row representing the cognate set. As can be seen from this Table, the assignment of all morphemes to a specific cognate set yields as many cognate sets as there are dialects. Given that quantitative approaches to phylogenetic reconstruction usually assume the development of all cognate sets to be independent, an assignment of all cognate morphemes to a single cognate set would therefore not only drastically increase the amount of cognate sets, but would also be entirely unrealistic, since these cognate morphemes surely did not evolve independently from each other.

In order to cope with the problems of indirect and partial cognacy, we decided to apply a very strict procedure of cognate assignment, grouping only those terms into cognate sets that correspond to identical sequences of Chinese characters. Since the data contained 244 entries for which no corresponding Chinese character was identified (and therefore no cognate assignment could be made), we excluded these entries. The remaining 9,957 words were grouped into 3,061 cognate sets. The cognate sets were then converted into a binary presence-absence matrix, where the columns represented the taxa, and the

TABLE 3 *Problem of partial cognacy in the Chinese dialects. The table shows cognate morphemes of translations of the concept 'moon' in seven Chinese dialects. As can be seen from the table, no two words are completely cognate, although all words share at least one cognate morpheme.*

	Cognate Sets						
Dialect	月	亮	光	呢	奶	明	爷
Shànghǎi	ɦyɪʔ$^{1-11}$	liã$^{13-23}$					
Wēnzhōu	nʮ$^{213-21}$		kuɔ^{33}				
Xiàmén	geʔ$^{5-21}$						
Jiàn'ōu	ŋyɛ^{42}			ni^{44}	nai^{33}		
Tàiyuán	yəʔ$^{2-54}$					mi^{45}	
Píngyáo	yʌʔ^{53}					mi$^{13-53}$	iɛ$^{13-31}$
Zhèngzhōu	yɛ^{24}				nai^{53}nai$^{53-24}$		

rows corresponded to distinct presence-absence patterns for a given cognate set, with 1 indicating the presence of a reflex and 0 indicating its absence. Since our method requires that a given cognate set has reflexes in at least two taxa, we excluded 2,005 cognate sets that were reflected only in one taxon. Our presence-absence matrix was thus reduced to a total of 1,056 presence-absence patterns.

2.3 *Reference Trees*

Our method estimates the extent to which the evolution of a set of characters (cognate sets reflected in the presence-absence patterns) can be explained by an evolutionary scenario that allows for only the vertical inheritance of characters. This scenario has to be defined with the help of a *reference tree* that captures the history of the language varieties under investigation. Given the specific sociolinguistic situation in China, the classification of the Chinese dialects is extremely difficult, and the opinions of scholars differ to a great extent (see Karlgren, 1954; Lǐ, 2005; Norman, 2003; Wáng, 2009, and the overview in Kurpaska, 2010: 36–62). The most common grouping distinguishes seven major dialect groups, namely (1) Mandarin (Guānhuà), (2) Xiāng, (3) Gàn, (4) Wú, (5) Hakka (Kèjiā), (6) Cantonese (Yuè), and (7) Mǐn (Norman, 1988: 181). However, alternative approaches that subdivide these varieties further are also quite popular, and at least three additional groups, namely Jìn (otherwise assigned to Mandarin), Huī (otherwise assigned to either Wú or Mandarin), and Pínghùa (otherwise assigned to Cantonese), are often proposed and discussed in the lit-

TABLE 4 *The dialect groups in our sample*

Group	Chinese	Altern. Grouping	# Dialects
Mandarin (Guānhuà)	管话		17
Jìn	晋	Mandarin	3
Xiāng	湘		2
Gàn	赣		1
Huī	徽	Wú, Mandarin	2
Wú	吴		4
Hakka (Kèjiā)	客家		2
Cantonese (Yuè)	粵		2
Pínghuà	平话	Cantonese (Yuè)	1
Mǐn	闽		6

erature (Kurpaska, 2010: 64–73). The ten major dialect groups are summarized in Table 4, along with alternative classifications and the number of varieties in our sample that belong to each group.

Most classifications group the Chinese dialects by comparing their deviation from the phonological system of Middle Chinese. One of the most salient features is the series of voiced plosives (*b, *d, *g, etc.) in Middle Chinese (Kurpaska, 2010: 35). These plosives show varying reflexes in the Chinese dialects. Sometimes they are retained completely (> b, d, g), sometimes all of them are devoiced (> p, t, k), sometimes the devoicing is accompanied by aspiration (> p^h, t^h, k^h), and sometimes the reflexes are split into a voiceless unaspirated and a voiceless aspirated series (> p/p^h, t/t^h, k/k^h). As Lǐ (2005) demonstrates, these reflexes are sufficient to distinguish six of the seven standard dialect groups, with Gàn and Hakka being merged into a single group.[4] However, the problem of this criterion (and most other classification criteria) is that they are merely used to *distinguish* certain dialect groups, while they do not *explain* how they developed. Although most classifications proposed thus far are based on historical criteria, few of them explicitly try to account for the genealogical development of the Chinese dialects.

4 Lǐ (2005) distinguishes different contexts in which the split of voiced to voiceless unaspirated and voiceless aspirated plosives occurred in order to distinguish Mǐn, Cantonese, and Mandarin.

Different theories have been proposed regarding the history of the major dialect groups. Among the most popular is Karlgren's (1954: 212) theory that almost all of today's Chinese dialects (except from the Mǐn dialects) go back to a *koiné* that was very widespread during the 6th century. He further states that this language was identical to Middle Chinese, the language whose phonological characteristics are recorded in the rime books that were compiled during that time. Norman (1988 and 2003) proposes a different theory, according to which Hakka, Cantonese, and Mǐn can be traced back to a common ancestor which split from the remaining dialects before the Middle Chinese period.

Based on these two different theories, we created two reference trees, one reflecting Norman's *Southern Chinese hypothesis*, and one reflecting Karlgren's *Common Chinese hypothesis*. In order to increase the distance between the trees, and since we could not determine the exact subgrouping of all major dialect groups from the literature alone, we added further differences to the subgroupings. Thus, in the Southern Chinese tree we grouped Wú and Huī dialects together, while in the Common Chinese tree we placed Huī closest to the Mandarin-Jìn group. In a similar way, we merged Hakka and Gàn in the Common Chinese tree following a reasonably popular proposal (see Sagart, 2002: 129–132), while assigning them to separate groups in the Southern Chinese tree. We also classified the Jìn dialects as a Northern Mandarin group in the Southern Chinese tree, while classifying them as first outgroup of Mandarin in the Common Chinese tree. For the internal subgrouping of the major dialect groups in both hypotheses, we generally employed the groupings proposed in the *Language Atlas of China* (Wurm and Liú, 1987). In cases where these groupings were too shallow and additional information was available, this internal subgrouping was further modified. Here, the internal classification of the Mǐn dialects was changed according to the classification in Norman (1991), and the eight groups of Mandarin dialects were further subdivided following Norman (1988).[5] Both reference trees for the major groups are given in Fig. 1. In order to test for possible differences between these 'traditional' reference trees and reference trees calculated from automatic approaches, we reconstructed two additional reference trees automatically. We applied the UPGMA algorithm (Sokal and Michener, 1958) and the Neighbor-joining

5 We are well aware of the fact that neither of the two trees can really claim to represent the true history of the Chinese dialects. However, as long as there are no detailed proposals regarding the genealogical classification of the Chinese dialects, we think it is more fruitful to accept uncertainties and possible mistakes resulting from the given trees than to abstain from the analysis in general.

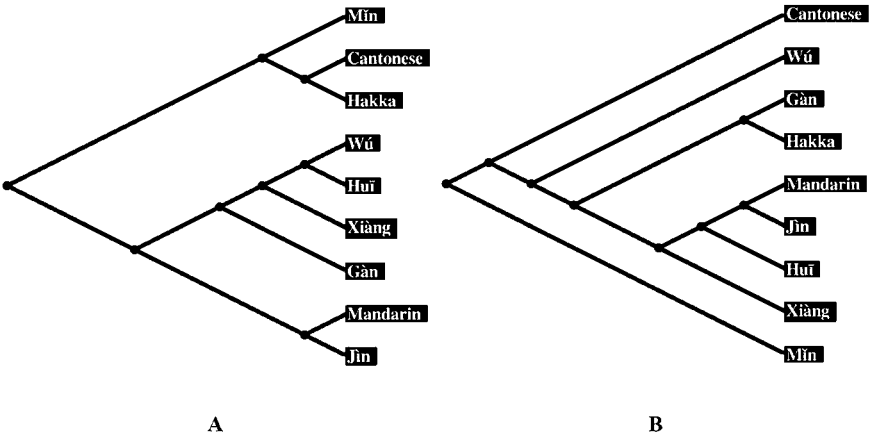

FIGURE 1 *Reference trees of the major groups for the Southern Chinese (A) and the Common Chinese (B) hypotheses. The reference trees are broadly based on the classifications of Norman (1988 and 2003) and Karlgren (1954), respectively, with the topologies expanded and adapted to accommodate the present sample (see text).*

algorithm (Saitou and Nei, 1987) to distance matrices derived from shared cognate percentages between all dialect pairs. The complete reference trees for all four analyses are given in Supplementary Material 1.

3 Methods

Building on the *minimal lateral network* (MLN) approach by Nelson-Sathi et al. (2011), our methods are based on an improved framework for the reconstruction of rooted phylogenetic networks (List et al., 2014). In contrast to the original approach, we introduce a refined method for *gain-loss mapping*. This method offers more flexible models with varying numbers of gain and loss events, captures multifurcation in reference trees, and also handles a certain amount of parallel evolution. Furthermore, we present a new method that derives *spatial networks* from rooted phylogenetic networks by plotting the results of the MLN approach to geographic maps. The new method is implemented as part of LingPy, an open source Python library for automatic tasks in historical linguistics (List and Moran, 2013, Version 2.2).

3.1 *Gain-Loss Mapping*

As pointed out before, any model of language evolution must take into account vertical as well as horizontal relations—i.e., borrowing. Borrowing processes

can be incredibly complex. Nevertheless, they usually leave observable traces, so that the borrowed word is often phonetically quite similar to the donor word. Furthermore, since the process of borrowing itself is not tree-like, borrowings that are mistaken for cognates can show up in form of presence-absence patterns that cannot be readily explained by the branching patterns of a family tree alone. As an example, compare the most widespread words for 'mountain' in the Germanic languages (German *Berg*, Dutch, Swedish *berg*, Danish *bjerg*) with the English word *mountain*. Assuming that English is a Germanic language, we see an astonishing difference to supposedly related languages. However, there is a striking similarity with words meaning 'mountain' in Romance languages such as Italian *montagna*, Spanish *montaña*, Portuguese *montanha*, and French *montagne*. If we had further evidence regarding the history of the languages and their branching patterns, there are two possible scenarios which could account for this coincidence: (1) English *mountain* is truly cognate with the Romance words, and reflexes of the word came to be lost in all other Germanic languages, or (2) English *mountain* was borrowed from one of the Romance languages, thereby replacing Old English *beorg*, the regular English reflex of Proto-Germanic *bergan 'mountain.' Given the branching pattern of the Germanic languages, it is much more plausible to assume the latter scenario (and indeed, historical evidence shows that English 'mountain' was borrowed from Old French *montaigne*). Thus, if languages show patterns of shared cognates that are in conflict with a given family tree, these patterns may be taken as a heuristic device for the detection of hitherto unrecognized borrowings.

As the example of English *mountain* shows, it is possible to gain some basic insights into language history by simply investigating the dynamics of gain and loss events. In evolutionary biology, the analysis of gain-loss scenarios (also called *presence-absence patterns* or *phyletic patterns*) is a common heuristic to identify possible instances of lateral gene transfer, and different methods for analyzing such patterns have been proposed in the recent past (see the overview in Cohen et al., 2010).

The basic idea of all these approaches is to create *gain-loss scenarios* for a given set of characters. A gain-loss scenario explains how a particular phyletic pattern could have evolved along a given reference tree. For a given pattern, each node of the tree is assigned to one of two possible states indicating the presence (1) or the absence (0) of the character in the pattern. *Events* are changes in the states from ancestral nodes to their direct descendants. A *gain event* (also called *origin*) is defined as the change from state 0 to state 1, and a *loss event* is defined as the change from state 1 to state 0. If the most appropriate analysis of a given phyletic pattern supports multiple gains (origins) of a character, this is usually taken as evidence for possible events of

TABLE 5 *Phyletic patterns of the cognate sets for 'mountain'*

Language	Spanish	Portuguese	French	English	German	Swedish
'mountain'	*montaña*	*montanha*	*montagne*	*mountain*	*Berg*	*berg*
Pattern M	1	1	1	1	0	0
Pattern B	0	0	0	0	1	1

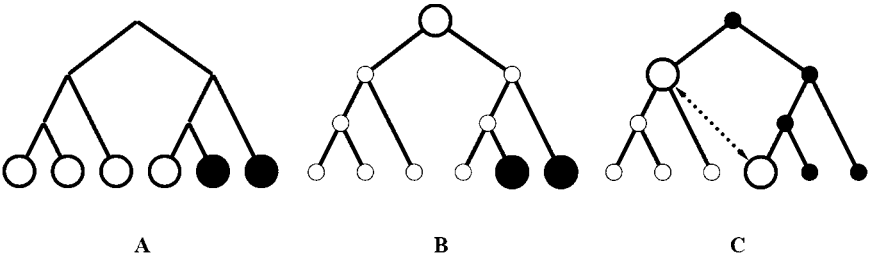

FIGURE 2 *Comparing alternative gain-loss scenarios. White nodes indicate the presence of a character, black nodes its absence. Large nodes indicate the respective event (gain or loss). In A, no scenario is inferred, B assumes one gain and two loss events, and C assumes two gain events and no loss event.*

lateral transfer (borrowing) that occurred during the evolution of the character. Table 5 illustrates how phyletic patterns are derived from the translation of 'mountain' into six Indo-European languages. For this group of languages, there are two different phyletic patterns, labeled M and B for convenience. Given the history of the six languages, Pattern B is unproblematic, supporting only a single origin hypothesis, with a loss of the character in English, and the gain of the character in the root. Pattern M (see Fig. 2A), however, can be mapped in two different ways: using a two-loss scenario as illustrated in Fig. 2B (scenario (1) above), or a two-gain scenario (scenario (2)), as illustrated in Fig. 2C. While the two-loss scenario infers that the character originated only once (in the root), the two-gain scenario infers two distinct origins for the character. Therefore, a lateral link between the two origins can be drawn, illustrated by the dotted line in Fig. 2C. This link is basically undirected, since it is not clear in which direction the borrowing event occurred. With this inference procedure, it is also not possible to determine when the link occurred, which explains why the link is drawn between the nodes in the tree where the characters originate.

Gain-loss scenarios can be inferred in different ways. Nelson-Sathi et al. (2011) follow Dagan and Martin (2007) in employing a *binary-branching top-*

down approach with different basic models, allowing for varying amounts of gains in a given phyletic pattern. The drawback of this approach is that the number of origins per phyletic pattern can only be an exponentiation of the base 2 (1, 2, 4, 8, 16, etc.), which results in a drastic restriction of the number of origins allowed by each model. A further drawback of this approach is that it can only be applied to bifurcating reference trees. This requirement is less problematic in biological applications since bifurcating reference trees are usually reconstructed automatically from the data. In linguistics, however, scholars are very cautious to propose detailed phylogenies, and multifurcating language trees (*soft polytomies* in the terms of Nunn, 2011: 22) are often used to reflect their uncertainty.

In order to overcome these shortcomings, we developed a *parsimony-based bottom-up approach* that allows for varying numbers of gains, depending on the phyletic pattern under investigation. In comparison with the top-down approach, our approach offers an increased number of models that can be tested on a given dataset. It also no longer restricts the maximal number of gain events that can be inferred by a given model, and—since the method is based on an exhaustive search of all possible scenarios—its application to multifurcating reference trees does not result in theoretical or practical problems.

Our approach is quite simple: given a phyletic pattern (a cognate set), there can be different gain-loss scenarios that could explain the evolution of the pattern. In order to find a consistent way of selecting the most parsimonious scenario, we test different *models* that assign different penalties for the scenarios, depending on the number of gain and loss events proposed by them. A model is defined as the ratio between penalties for gain and loss events. The model 2–1, for example, penalizes gain events with 2 and loss events with 1. The most parsimonious scenario for a given model is the one which minimizes the overall penalty. In order to compute all possible gain-loss scenarios, we use a bottom-up approach that starts from the leaves and climbs up to the root, thereby storing all different possibilities of character evolution. Basically, our approach is brute-force.

The search space can, however, be efficiently restricted. Firstly, when climbing up the reference tree in order to calculate the possible scenarios, we can exclude those which exceed the *maximum number of gain events* allowed on each path from the root of the tree to its leaves. If this number is set to 1 (as it is by default in our approach), this means that, on a given path, characters cannot be gained, lost, and gained again. This is a simplifying requirement, since it is possible that characters on a given lineage are lost and afterwards reintroduced as borrowings—an example being English 'flower,' which was borrowed from Old French *flour* which goes back in turn to Latin *flōre*(*m*). The Latin word is

cognate with English *blossom* and German *Blume* 'flower,' all being reflexes of Proto-Indo-European *b^hleh_3- 'blossom' (de Vaan, 2008: 227). A strict modeling of the complicated history of these words with help of gain-loss scenarios would require us to assume that the character was lost and gained again in English. However, given that these cases are very rare, allowing for them would not only bloat our search space, but also affect the results in a way that is difficult to control.

Secondly, having determined the scenarios that do not exceed our *maximum gain criterion*, we can filter them further by storing only those scenarios with minimal weight. Here, it is important to keep in mind that a scenario with minimal weight on a given subtree is not necessarily a scenario with minimal weight in general. Since, when climbing the reference tree, one cannot tell whether the character state of the temporary root node is an event (a change of the character state) or not, it is possible that a given scenario seems to be cheap at a certain point in the calculation but later turns out to be much more expensive. In order to prevent the model from missing good scenarios, we carry out a separate filtering of those scenarios in which the character in the temporary root node is present and those in which it is absent. Since unpredictable costs of subtree scenarios depend only on the state of the temporary root character, this guarantees that our approach always finds the most parsimonious scenario. It is possible that there is more than one scenario that minimizes the penalty. In such a case we first select the scenario with the minimal amount of gain events, and if there is still more than one scenario, we follow the proposal by Mirkin et al. (2003) and select the scenario in which the gain events are closest to the leaves of the reference tree.

As an example, compare the two-loss scenario in Fig. 2B with the two-gain scenario in Fig. 2C. For the two-loss scenario, the 2–1 model yields a total score of 4 ($1\times2+2\times1$), since there are two losses and one gain.[6] The two-gain scenario in Fig. 2C also yields a score of 4 ($2\times2+0\times1$). In this case, we choose the model which infers the minimal amount of gains, and the two-loss model is chosen as the most parsimonious one. Changing the model to 1–1 yields penalties of 3 ($1\times1+2\times1$) for the two-loss scenario and 2 ($2\times1+0\times1$) for the two-gain scenario. In this case, the two-gain scenario is the most parsimonious.

6 We follow Mirkin et al. (2003) in counting the presence of a character in the root as a normal gain event.

3.2 *Finding Optimal Gain-Loss Models*

Gain-loss mapping is useful for testing possible scenarios of character evolution. However, as long as there is no direct criterion that helps to choose the best of many solutions, the method hardly gives us any new insights. Here, we follow Nelson-Sathi et al. (2011) in using the distribution of *ancestral vocabulary sizes* as a criterion to determine the best model for a given dataset. The basic idea behind this criterion for model selection is that the number of words that ancestral languages use to express a given set of concepts should not differ greatly from the number of words used by the contemporary languages. When assuming that English *mountain* is not a borrowing but a retention (two-loss scenario), this would force us to trace the word back to Proto-Germanic. However, since the counterparts of 'mountain' in the rest of the Germanic languages also point to a common origin, this would necessitate the assumption that there were two words denoting the concept 'mountain' in Proto-Germanic. Although multiple synonyms for a given concept are not impossible, they are rather unlikely to occur frequently; and since our approach is applied to large datasets and not to single items, it seems reasonable to assume that a model explaining the given data adequately should be preferred to a model that yields much larger amounts of synonyms in the ancestral languages than are attested in the contemporary ones. In the case of *mountain*, this means that the 1–1 model should be preferred to the 2–1 model, since the latter favors the two-loss scenario and thus entails the assumption of more synonyms in the ancestral languages.

One could argue that the growing amounts of synonyms in ancestral languages can be explained by assuming the words had different meanings in those languages. English *mountain*, for example, could be derived from Proto-Indo-European *mon-ti 'protrusion, height,' which is the presumed ancestor of Latin *mōns* (de Vaan, 2008: 388). Such a scenario, however, is rather unlikely, since it presupposes that the same semantic shift from 'height' to 'mountain' occurred in the Romance languages and in English. While parallel semantic shift is not improbable *per se*, it is rather unlikely when involving the *same* source forms in *independent* branches of a language family. Furthermore, even if it was frequent, it would not disfavor vocabulary size distributions as a criterion for model selection. It would merely change what gain-loss mapping techniques can infer.

In order to compare how well a given model accounts for the vocabulary size criterion, we compute the number of characters present in the ancestral nodes of the reference tree by tracing all origins inferred by the model back to the respective nodes. We then use the Wilcoxon rank-sum test (see the description in Kruskal, 1957) to test the hypothesis that the ancestral and the contemporary

TABLE 6 *Patchy cognate sets for 'mountain.' In contrast to the cognate set in Table 5, pattern* M *is now split into two distinct patterns:* M_1 *and* M_2.

Language	Spanish	Portuguese	French	English	German	Swedish
'mountain'	*montaña*	*montanha*	*montagne*	*mountain*	*Berg*	*berg*
Pattern M_1	1	1	1	0	0	0
Pattern M_2	0	0	0	1	0	0
Pattern B	0	0	0	0	1	1

vocabulary distributions are likely to be drawn from the same sample. Since we cannot exclude the possibility that parallel evolution influences our results, we modified our method in such a way that it allows for a certain amount of parallel evolution. This can be done in a very straightforward way by using a scaling factor to decrease the ancestral vocabulary sizes before the Wilcoxon rank-sum test is applied. As a default, this scaling factor is set to 5%. Thus, we allow ancestral vocabulary size distributions to grow up to 5% larger than contemporary ones.

Having determined a model that explains the phyletic patterns of a given dataset in such a way that the distribution of ancestral and contemporary vocabulary sizes does not differ significantly, the results of the analysis can then be displayed by splitting all cognate sets for which more than one origin was inferred into secondary subsets, as illustrated in Table 6. These *patchy cognate sets* (PCS) can then be further analyzed in different ways. One could, for example, compare the correctness of the original cognate assignments by checking the sound correspondences between the distinct subsets for irregular patterns. In the case of English *mountain*, there is an irregular correspondence between the English [t] and the [t] in the Romance languages, where we would expect a [d] if it were a regular correspondence (compare English tooth [tuːθ] vs. French *dent* [dɑ̃] 'tooth').

3.3 *Minimal Lateral Network*

Another way to analyze the results further is to reconstruct a *minimal lateral network* (MLN) from the inferred gain-loss scenarios (Dagan et al., 2008; Nelson-Sathi et al., 2011). The MLN is a weighted network that displays patterns of vertical and lateral inheritance. The reference tree is used to represent patterns of vertical inheritance between the contemporary and the ancestral languages. Additional edges drawn between the nodes of the reference tree represent possible borrowing events. Borrowing events are assumed for all patterns for which

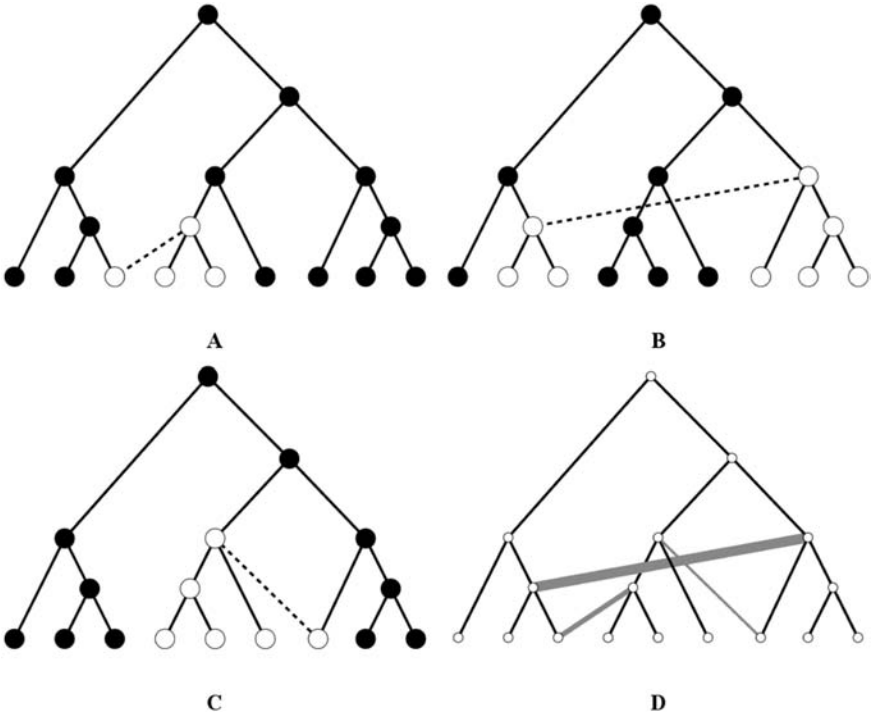

FIGURE 3 *Minimal lateral network reconstruction. If more than one origin is inferred for a given phyletic pattern, the nodes where the characters originate are connected by lateral edges (A–C). In the MLN (D), the edges inferred for all patterns are combined, with edge weights (visualized as differences in line width) reflecting the number of occurrences.*

more than one origin was inferred by a given gain-loss model, and links are drawn between the nodes in which the characters originate. The weights of these lateral edges reflect the number of patterns that support a given link. Figure 3 illustrates this procedure. In Figs 3A–C, three different links are drawn between nodes from which different characters originate more than once on the reference tree. If the number of patterns supporting these scenarios in a given dataset differs, with Fig. 3A occurring twice, Fig. 3B four times, and Fig. 3C once, we arrive at a weighted network for the whole dataset as shown in Fig. 3D.

Drawing lateral links between characters that originate from two different nodes is easy, since there is only one link that can be drawn to connect them. However, if a gain-loss scenario yields more than two separate origins for a given character, there are as many as $(n^2–n)/2$ possible edges which can be drawn to connect n nodes. While drawing all possible edges would surely cover all possibilities, it is obviously unrealistic: since borrowing is a directed

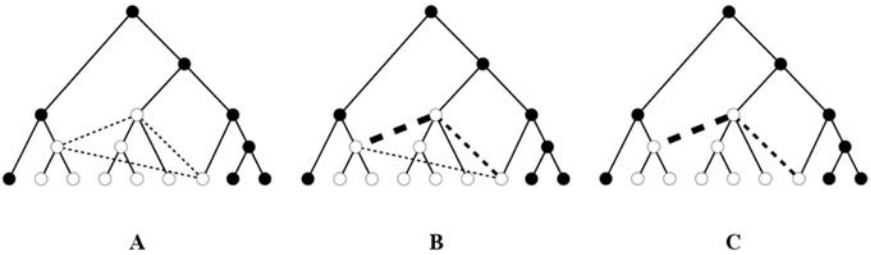

FIGURE 4 *Removing redundant lateral edges in the minimal lateral network. A shows the initial stage. B shows the intermediate stage after edge weights have been inferred for all lateral edges. C shows the resulting minimum spanning tree.*

process that involves a donor and a recipient language, such a scenario would indicate that all languages are both donors and recipients. In order to solve this problem, the complete graph representing all hypothetical connections has to be reduced to a graph consisting of n–1 edges that connects all nodes (a spanning tree). Given that, according to Cayley's (1889) formula, a complete graph of n edges has as many as n^{n-2} spanning trees, it is important to apply a consistent criterion to select one of these trees. The most straightforward way to do so is to select a *minimum spanning tree*, that is, a tree that minimizes the sum of the edge weights.[7] For gain-loss scenarios involving more than two origins, we determine the edge weights for all node pairs *ni* and *nj* by calculating the number of shared multiple origins of *ni* and *nj* in all phyletic patterns of the data. We then convert these weights to distances and use Kruskal's (1956) algorithm to calculate the minimum spanning tree between the nodes. This procedure is illustrated in Fig. 4. This is equivalent to assuming that potential donor lineages with a high frequency of occurrence in the sample have a higher probability of donating than low-frequency potential donor lineages.

3.4 *Minimal Spatial Network*

A minimal lateral network is useful to evaluate the degree to which the evolution of a set of characters follows the presumed branching pattern of a set of languages. However, since languages are not only spoken at a specific time, but also in a specific place, it seems useful to plot the inferred lateral connections onto a geographic map. This may be helpful both for evaluating the results of a given analysis and getting an impression of major diffusion areas. When

7 In our case it would be more appropriate to call it a 'maximum spanning tree,' since the edge weights in the MLN do not represent distances but similarities between nodes.

reconstructing a *minimal spatial network* (MSN) from a given MLN, only the leaves can be plotted because the ancestral nodes have few geographical constraints, so their inclusion in the graph would add too much cluttering information. Therefore the internal nodes of the MLN (the ancestral taxa) are removed and, as a result, internal edges (edges between contemporary and ancestral taxa, and edges between ancestral taxa) are lost. In order to retain information that is connected within them when constructing the spatial network, we project information from internal nodes onto leaves. As a selection criterion to link information from internal to external nodes, we use a simplified approach based on geographic distance. If an edge originally connects an internal node *ni* and an external node *ne*, we first determine all descendant nodes of *ni* on our geographic map. We then draw a convex hull around all descendant nodes of of *ni* and connect the descendant node of *ni* that is (a) on the hull and (b) geographically closest to our external node *ne*. For two internal nodes, we proceed in a similar way, the difference being that two convex hulls are drawn around the descendants of the two internal nodes, and the two geographically closest nodes which appear on the hulls are connected. The central idea behind this approach is that ancestral languages can be represented by the area covered by their descendants.

4 Results

4.1 *Gain-Loss Models for Southern and Common Chinese*

We applied our analysis to the Southern Chinese, the Common Chinese, and the two automatically reconstructed reference trees, using five different gain-loss models with varying penalties for gains and losses: 3–1, 5–2, 2–1, 3–2, and 1–1. We then compared the resulting distributions of ancestral and contemporary vocabulary sizes in order to determine which of the models would fit the data best. For all reference trees, there are two gain-loss models (5–2 and 2–1) in which the vocabulary size distributions do not differ significantly ($\alpha = 0.05$). In all cases, the 2–1 model is the one with the highest probability ($p = 0.73$ for Southern Chinese, $p = 0.76$ for Common Chinese, $p = 0.84$ for UPGMA, and $p = 0.55$ for Neighbor-joining).[8]

As far as the gain-loss models are concerned, the differences between the four trees do not seem to alter gain-loss mapping analyses greatly. Basically,

8 A comparison of the vocabulary size distributions inferred for all analyses is provided in Supplementary Material II.

TABLE 7 *Basic results of the analyses*

Comparandum	Southern Chinese	Common Chinese	Neighbor-joining	UPGMA
Best model	2–1	2–1	2–1	2–1
p-value	0.73	0.76	0.55	0.84
Patchy cognates	567 (54%)	557 (53%)	510 (48%)	585 (55%)
Average n. of origins	1.97	1.81	1.81	2.00
Maximal n. of origins	9	9	8	8

this also holds for some further general characteristics of the models, such as the average number of origins, the number of patchy cognate sets, and the maximum number of origins, all of which are displayed in Table 7.[9] The Neighbor-joining reference tree outperforms the other trees by yielding the lowest percentage of patchy cognate sets. However, since the Neighbor-joining tree itself is in conflict with traditional dialect classification, this merely shows that the Neighbor-joining method is good in maximizing the tree-like signal in the data. It does not mean that the results are necessarily more realistic. Comparing these results with those of List et al. (2014) for Indo-European languages, it is interesting to note that the percentage of patchy cognate sets is quite different (48–55% for the Chinese analyses, but 31% for Indo-European). Given the complex history of the Chinese dialects, this is not surprising but, rather, in agreement with our expectations.

4.2 *MLN and MSN*

Having determined a gain-loss model that brings ancestral and contemporary vocabulary size distributions closely together, we can use this scenario to reconstruct a minimal lateral network. Figure 5 shows the MLN reconstructed for the Southern Chinese reference tree. Interestingly, the heaviest edges occur inside the Mandarin and the Jìn dialects. Here, the Zhèngzhōu dialect plays a central role, having a remarkable number of connections not only to the ancestral node

9 Note that in Table 7 and throughout this paper, the term 'origins' refers to events that distribute a given cognate across dialects and geographical ranges. Thus, inferring 8 or 9 origins in Table 7 does not suggest 8 or 9 independent origins, it simply means that 8 or 9 events are inferred, under our minimizing premises, to underlie its current geographical and dialectic distribution.

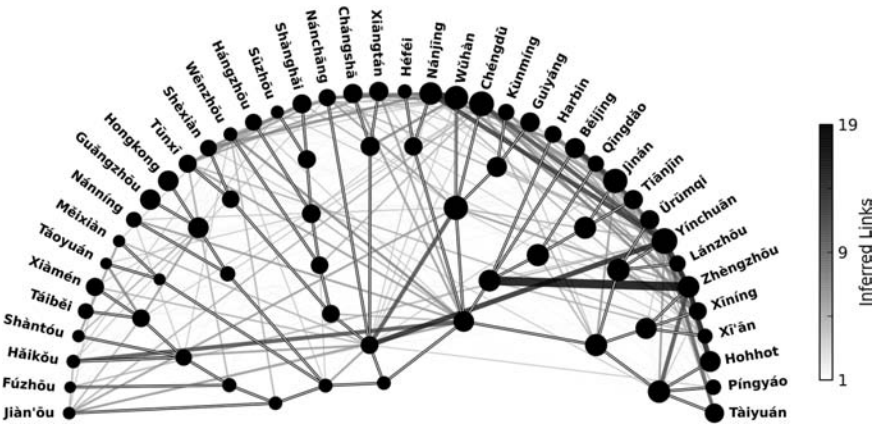

FIGURE 5 *The minimal lateral network of the Southern Chinese reference tree. The node size reflects the inferred number of cognate sets in each language variety. The links reflect the minimal number of lateral transfer events that is required to minimize the differences between the ancestral and the contemporary vocabulary size distribution.*

of the Northern Mandarin dialects (19 shared patchy cognate sets, PCSs), but also to Lányín Mandarin (11 PCSs with Yínchuān), and Jìn (11 links with Hohhot, 9 links with the ancestral node of Jìn). The fact that Zhèngzhōu is not grouped with the Zhōngyuán Mandarin dialects in both automatic analyses (see Supplementary Material I) further reflects the uncertain status of this dialect. Apart from the central role that Zhèngzhōu plays in the Southern Chinese MLN, there is a remarkable number of inferred connections between the Jìn dialects and the Northern and Northwestern Mandarin dialects. Both the role of Zhèngzhōu and the multitude of links between Jìn dialects and Northern and Northwestern Mandarin can also be reported for the Common Chinese analysis (see Supplementary Material III). The status of the Jìn dialects as a group separate from Mandarin is highly disputed in Chinese dialectology (Kurpaska, 2010: 74 f.). If their separation is justified, our method shows that they are highly influenced by neighboring varieties.

The heavy links between Northern and Northwestern Mandarin and Jìn dialects can be more easily recognized in the minimal spatial network shown in Fig. 6. Apart from the high and also quite unexpected diversity in the north, one can find interesting connections in the south-east, where the greatest number of generally recognized distinct dialect groups is found. Thus, Xiāngtán and Chángshā, the two Xiāng dialects in our sample, show their strongest connections to neighboring Mandarin dialects. That the Xiāng dialects have undergone a strong influence from Mandarin dialects has been noticed in the literature for a long time (Norman, 1988: 207 f.). Even more interesting is the strong

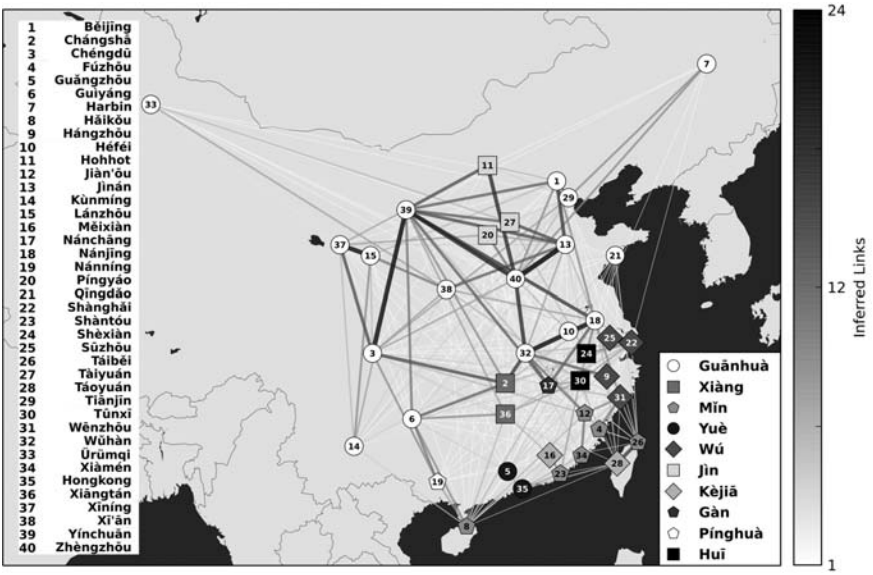

FIGURE 6 *The minimal spatial network of the Southern Chinese reference tree. The links reflect the external and the internal edges between all contemporary language varieties as inferred in the minimal lateral network.*

link between the Wú dialect Wēnzhōu and its neighboring Mǐn dialects.[10] This link is surprising, since in Chinese dialectology it is usually assumed that the border between the Mǐn and the Wú dialects is rather strict (ibid.: 189). However, a closer inspection of the words in Wēnzhōu that are patchily distributed shows that it is indeed very likely that they have been borrowed from the Mǐn dialects, since they are not found in the other Wú dialects, but are quite representative of the Mǐn varieties. Thus, among others, we find that the Wēnzhōu word for 'chopsticks' is [dzei^{22}] with the corresponding character 箸. This is a very archaic expression for 'chopsticks' that is almost exclusively reflected in the Mǐn dialect area. Most other dialects (including all other Wú dialects in our sample) have replaced it with cognate forms of Common Chinese *kuàizi* 筷子 (see Norman, 1988: 76 for details regarding the origin of *kuàizi*). Similar examples where Wēnzhōu has a form that is not reflected in the other Wú dialects, but common in the Mǐn dialects include:

10 In the MSN, the link is drawn between Wēnzhōu and Jiàn'ōu. This is, however, an artifact of the spatial representation. In the underlying MLN, the link is between Wēnzhōu and several ancestral nodes of the Mǐn dialects.

Wēnzhōu [dɤu^{31}] 头 'classifier (for cows and pigs)' (compare Shànghǎi [tsaʔ^{5}] 只), Wēnzhōu [tɕaŋ^{33}ko^{33}] 金瓜 'pumpkin' (compare Shànghǎi [ve^{22}ko^{44}] 南瓜), and Wēnzhōu [liɛ^{35}bu^{13}] 龙雹 'hail' (compare Shànghǎi [piŋ^{55}bɔ^{21}] 冰雹).[11]

Above we have seen that differences in the reference trees did not affect the gain-loss models. This was also observed in Nelson-Sathi et al.'s (2011) analysis of the Indo-European languages and is indicative of a high level of patchiness in the cognate distribution—for data with a comparatively large component of non-treelike structure, the influence of the reference tree becomes less crucial. What was also noted in the study of Nelson-Sathi et al. (2011), however, is that changes in the reference tree may have an impact on the concrete predictions of a given model, indicating in turn that there are detectable vertical components in the data. For our two reference trees in the present Chinese dataset, we can report similar findings. Although the agreement between the Southern and the Common Chinese analyses regarding the detection of patchy cognate sets is rather high, with 966 out of 1056 cognate sets (91%) being identically identified as either patchy or non-patchy cognate sets, many differences in the specific individual scenarios are still observable. Table 8 gives unweighted and weighted degrees for the five most connective nodes in the MLNs for Southern Chinese and Common Chinese.[12] Although four of the five most connected nodes appear in both analyses, they differ greatly regarding their unweighted and their weighted degrees. Since we do not know which of the two scenarios reflects the historical process more closely, we are currently limited to noting the differences. In future studies, it could be of interest to identify independent criteria by which to compare the probabilities of different weighted degrees for given (sets of) nodes, and to use these criteria to evaluate the attributes of different reference trees.

4.3 *Influence of Standard Chinese*

One point we have not addressed so far is the role of the *Dachsprache* in our data. Given that Standard Chinese derived from the dialect of Běijīng, it is surprising that this dialect only plays a minor role in the networks shown in Figs 5 and 6. Běijīng does not appear among the top five nodes with the highest

11 A full account of all the inferred patchy cognates for the Southern Chinese analysis is given in Supplementary Material IV.

12 The degree is the number of edges connecting to a given node in a graph. The weighted degree is calculated by summing up the weights for all edges of a given node (cf. Newman, 2004).

TABLE 8 *Comparing the nodes with the highest degrees for the Southern Chinese* (A) *and the Common Chinese analysis* (B)

Taxon	Degree		Taxon	Degree	
	Unweighted	Weighted		Unweighted	Weighted
Nánjīng	29	81	Shèxiàn	27	58
Zhèngzhōu	29	105	Chéngdū	26	69
Yínchuān	27	114	Yínchuān	26	93
Chéngdū	26	72	Jìnán	24	58
Jìnán	26	70	Nánjīng	24	70
A			B		

degrees (either unweighted or weighted), nor is it involved in any of the heaviest edges. The fact that Běijīng and Standard Chinese played a less pronounced role than expected might be due to a certain shortcoming in our method. Gain-loss mapping requires that borrowing events are still detectable due to patterns that cannot be explained by a reference tree. Borrowing, however, can become so frequent that patchy distributions are no longer detectable.[13] If a word is borrowed (or is actively introduced) by all taxa of a given branch so that the existence of its predecessors is masked, the gain-loss mapping approach assumes that these words are all inherited from a common ancestor language and so no patchy distributions are detected. If, however, the ancestral words have not died out and still exist in refugia that can be detected through more thorough geographical sampling, these effects should be detectable and, in principle, quantifiable.

Although the networks themselves do not give us a hint, the influence of Standard Chinese on Chinese dialect history can still be identified when comparing how many of the cognate sets in each dialect are actually patchy. In Table 9, the five dialects that show the largest frequencies of patchy cognate sets per number of words are listed. In this list, the Běijīng dialect as the closest representative of Standard Chinese occupies the first position, showing the highest ratio of patchy cognate sets per word in both the Southern Chinese

13 In genetics, there is the term 'selfish DNA' to describe genes that can rapidly increase their frequency through spread, because they are readily able to spread (*transposons*). There is also the concept of positive selection, which can lead to the very rapid spread and fixation of new alleles in a population.

TABLE 9 *Comparing the average number of patchy cognates per dialect in the Southern Chinese* (A) *and the Common Chinese analysis* (B)

Taxon	# Words	PCS	Ø	Taxon	# Words	PCS	Ø
Běijīng	236	95	0.40	Běijīng	236	99	0.42
Zhèngzhōu	278	108	0.39	Chéngdū	320	127	0.40
Tiānjīn	253	97	0.38	Zhèngzhōu	278	110	0.40
Jìnán	315	120	0.38	Tiānjīn	253	100	0.40
Chéngdū	320	121	0.38	Nánjīng	276	107	0.39
A				B			

and the Common Chinese analysis. This shows that Běijīng and Standard Chinese play a definite role in our network, although this role is currently not quantifiable in terms of degree and heavily weighted edges, but only in the patchy cognate sets themselves.

5 Discussion

In evolutionary biology and historical linguistics, the term *phylogenetic network* is often used in a very broad sense, referring to 'any graph used to represent evolutionary relationships (either abstractly or explicitly) between a set of taxa that labels some of its nodes (usually the leaves)' (Huson et al., 2010: 69). Given the fuzziness of this definition, Morrison (2011: 42) suggests drawing a further distinction between two types of phylogenetic networks: *data-display networks* and *evolutionary networks*. Data-display networks are merely a data summary, while evolutionary networks represent a direct phylogenetic hypothesis which 'should display evolutionary relationships between ancestors and descendants' (ibid.: 43). According to this definition, the popular *split networks* (Huson et al., 2010: 71f.), which were also applied to Chinese dialect data (Ben Hamed and Wang, 2006), are data-display networks; the networks we reconstructed with our method come close to evolutionary networks, since they display both patterns of vertical and lateral inheritance. Nevertheless, while our method appears to be pointing in the right direction with regard to uncovering vertically and horizontally shared components in phylogenetic analyses, it is clear that there are still many problems that need to be addressed in future studies.

Our method relies heavily on the accuracy of proposed assessments of etymological relatedness. If the data is incorrectly coded, the results will be off

the mark, but that is true of any analytic method, not just networks. The fact that differences regarding homology judgments can have a great impact on the results reported for gene distributions across genomes was shown in a study by Dagan and Martin (2007: 873), where varying sizes of gene families had a deeper impact on gain-loss models and estimated rates of lateral gene transfer than differences in reference trees. Our current approach to conducting cognate judgments is very strict. Even the slightest morphological variation that might result from regular processes of affixation will force us to separate words into different cognate sets. Although we think that the requirement of direct cognacy as opposed to partial or oblique cognacy is a necessary and reasonable requirement for our method, we recognize that the borders can overlap. Furthermore, it is highly likely that we missed many cases of valid, direct cognacy by conducting cognate judgments on the basis of the identity of the Chinese character sequences. This is a parameter that can be varied in future analyses.

The fact that our networks alone did not uncover the influence of Standard Chinese, and that its influence could only be shown when comparing the number of patchy cognate sets per number of words in a given variety, points to a general problem of the current method for network reconstruction. At the moment, our method simply connects those nodes on the reference tree for which a patchy cognate set has been inferred by a given gain-loss model. In this sense, our approach is greedy. The specific borrowing process, however, cannot be inferred with the method, since it neither points to a direction of the process, nor does it point to a concrete source, since in many cases the gain-loss model infers that characters originate on internal (ancestor) rather than external (contemporary) nodes. Although our method is an improvement over data-display networks, it is still an effort to translate its results into inferred historical processes.

Despite these drawbacks, we are confident that it is worthwhile to pursue this road further. Borrowing is an integral component of language history and the networks can accommodate this mechanism in a way that no bifurcating tree can. Our method clearly shows that the tree model also fails to explain the majority of the lexical data of the Chinese dialects in our sample. Not only does it confirm general uncertainties of Chinese dialect classification that have been long discussed, it also reveals the strong influence of the standard language on the diatopic varieties of Chinese, uncovering a small sketch of the complexity of Chinese dialect history.

Supplementary Material and Software

The Supplementary Material accompanying this study contains figures of all reference trees that were used for this study (Supplementary Material I), the vocabulary size distributions inferred for all analyses (Supplementary Material II), the MLN and MSN for the Common Chinese analysis (Supplementary Material III), and a full account of all patchy cognate sets inferred for the Southern Chinese analysis (Supplementary Material IV). The materials can be downloaded from:

http://www.molevol.de/resources/index.html?id=011list2014/

A Python script that replicates the analyses upon which this study was based along with the input data is available under:

https://gist.github.com/LinguList/7481097.

Acknowledgments

This research was supported by the *German Federal Ministry of Education and Research* (BMBF, research project 'Evolution and Classification in Biology, Linguistics, and the history of the Sciences' http://www.evoclass.de) and the ERC grants 240816 and F020515005.

References

Atkinson, Quentin D. and Russell D. Gray. 2006. How old is the Indo-European language family? Illumination or more moths to the flame? In Peter Forster and Colin Renfrew (eds.), *Phylogenetic Methods and the Prehistory of Languages*, 91–109. Cambridge: McDonald Institute for Archaeological Research.

Barbour, Stephen and Patrick Stevenson. 1998. *Variation im Deutschen. Soziolinguistische Perspektiven*. Berlin: de Gruyter.

Baxter, William H. 1992. *A Handbook of Old Chinese Phonology*. Berlin: Mouton de Gruyter.

Ben Hamed, Mahé and Feng Wang. 2006. Stuck in the forest: Trees, networks and Chinese dialects. *Diachronica* 23: 29–60(32).

Bonfante, Giuliano. 1931. I dialetti indoeuropei. *Annali del R. Istituto Orientale di Napoli* 4: 69–185.

Branner, David Prager. 2006. Some composite phonological systems in Chinese. In David Prager Branner (ed.), *The Chinese Rime Tables. Linguistic Philosophy and Historical-comparative Phonology*, 209–232. Amsterdam: Benjamins.

Cayley, Arthur. 1889. A theorem on trees. *Quarterly Journal of Pure and Applied Mathematics* 13: 376–378.

Cohen, Ofir, Haim Ashkenazy, Frida Belinky, Dorothée Huchon, and Tal Pupko. 2010. GLOOME: Gain loss mapping engine. *Bioinformatics* 26.22: 2914–2915.

Coseriu, Eugenio. 1973. *Probleme der strukturellen Semantik. Vorlesung gehalten im Wintersemester 1965/66 an der Universität Tübingen*. Tübingen: Narr.

Dagan, Tal, Yael Artzy-Randrup, and William Martin. 2008. Modular networks and cumulative impact of lateral transfer in prokaryote genome evolution. *Proceedings of the National Academy of Sciences of the U.S.A.*. 105.29: 10039–10044.

Dagan, Tal and William Martin. 2007. Ancestral genome sizes specify the minimum rate of lateral gene transfer during prokaryote evolution. *Proceedings of the National Academy of Sciences of the U.S.A.*. 104.3: 870–875.

Goossens, Jan. 1973. *Niederdeutsch. Sprache und Literatur. Eine Einführung*. Neumünster: Karl Wachholtz.

Gray, Russell D. and Quentin D. Atkinson. 2003. Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature* 426.6965: 435–439.

Hóu Jīng 侯精 (ed.). 2004. *Xiàndài Hànyǔ fāngyán yīnkù* 现代汉语方言音库 [Phonological database of Chinese dialects]. Shànghǎi 上海: Shànghǎi Jiàoyù 上海教育.

Huson, Daniel H., Regula Rupp, and Celine Scornavacca. 2010. *Phylogenetic Networks. Concepts, Algorithms, and Applications*. Cambridge: Cambridge University Press.

Karlgren, Bernhard. 1954. Compendium of phonetics in ancient and archaic Chinese. *Bulletin of the Museum of Far Eastern Antiquities* 26: 211–367.

Koonin, Eugene V. 2005. Orthologs, paralogs, and evolutionary genomics. *Annual Review of Genetics* 39: 309–338.

Kruskal, Joseph B. 1956. On the shortest spanning subtree of a graph and the traveling salesman problem. *Proceedings of the American Mathematical Society* 7.1: 48–50.

Kruskal, William H. 1957. Historical notes on the Wilcoxon unpaired two-sample test. *Journal of the American Statistical Association* 52.279: 356–360.

Kurpaska, Maria. 2010. *Chinese Language(s). A Look through the Prism of The Great Dictionary of Modern Chinese Dialects*. Berlin and New York: de Gruyter.

Lǐ Xiǎofán 李小凡. 2005. *Hànyǔ fāngyán fēnqū fāngfǎ zài rènshi* 汉语方言分区方法再认识 [Reevaluating the classification of the Chinese dialects]. *Fāngyán* 方言 4: 356–363.

List, Johann-Mattis and Steven Moran. 2013. An open source toolkit for quantitative historical linguistics. In *51th Annual Meeting of the Association for Computational Linguistics (ACL 2013), Proceedings of the Conference System Demonstrations, Aug. 4–9, 2013, Sofia, Bulgaria*, 13–18. Stroudsburg, PA: Association for Computational Linguistics.

List, Johann-Mattis, Shijulal Nelson-Sathi, Hans Geisler, and William Martin. 2014.

Networks of lexical borrowing and lateral gene transfer in language and genome evolution. *BioEssays* 36.2: 141–150.

McMahon, April and Robert McMahon. 2005. *Language Classification by Numbers*. Oxford: Oxford University Press.

Meiser, Gerhard. 1998. *Historische Laut- und Formenlehre der lateinischen Sprache*. Darmstadt: Wissenschaftliche Buchgesellschaft.

Mirkin, Boris G., Trevor I. Fenner, Michael Y. Galperin, and Eugene V. Koonin. 2003. Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. *BMC Evolutionary Biology* 3: 2 (doi: 10.1186/1471-2148-3-2).

Morrison, David A. 2011. *An Introduction to Phylogenetic Networks*. Uppsala: RJR Productions.

Nelson-Sathi, Shijulal, Johann-Mattis List, Hans Geisler, Heiner Fangerau, Russell D. Gray, William Martin, and Tal Dagan. 2011. Networks uncover hidden lexical borrowing in Indo-European language evolution. *Proceedings of the Royal Society B. Biological Sciences* 278.1713: 1794–1803.

Newman, Mark E.J. 2004. Analysis of weighted networks. *Physical Review E* 70.5: 056131.

Norman, Jerry. 1988. *Chinese*. Cambridge: Cambridge University Press.

Norman, Jerry. 1991. The Mǐn dialects in historical perspective. In William S.-Y. Wang (ed.), *Languages and Dialects of China* (Journal of Chinese Linguistics, Monograph Series Number 3), 325–360. Berkeley: Project on Linguistic Analysis.

Norman, Jerry. 2003. The Chinese dialects: Phonology. In Graham Thurgood and Randy LaPolla (eds.), *The Sino-Tibetan Languages*, 72–83. London: Routledge.

Nunn, Charles L. 2011. *The Comparative Approach in Evolutionary Anthropology and Biology*. Chicago: University of Chicago Press.

Sagart, Laurent. 2002. Gan, Hakka and the formation of Chinese dialects. In Dah-an Ho (ed.), *Dialect Variations in Chinese. Papers from the Third International Conference on Sinology, Linguistics Section*, 129–153. Taipei: Academia Sinica.

Saitou, Naruya and Masatoshi Nei. 1987. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* 4.4: 406–425.

Schleicher, August. 1853. Die ersten Spaltungen des indogermanischen Urvolkes. *Allgemeine Monatsschrift für Wissenschaft und Literatur*: 786–787.

Schmidt, Johannes. 1872. *Die Verwandtschaftsverhältnisse der indogermanischen Sprachen*. Weimar: Hermann Böhlau.

Schuchardt, Hugo. 1900. *Über die Klassifikation der romanischen Mundarten. Probe-Vorlesung gehalten zu Leipzig am 30. April 1870*. Graz: Styria. Downloadable at http://schuchardt.uni-graz.at/werk/pdf/309 (accessed June 4, 2014).

Sokal, Robert R. and Michener, Charles D. 1958. A statistical method for evaluating systematic relationships. *University of Kansas Scientific Bulletin* 28: 1409–1438.

Southworth, Franklin C. 1964. Family-tree diagrams. *Language* 40.4: 557–565.

Swadesh, Morris. 1952. Lexico-statistic dating of prehistoric ethnic contacts. With special reference to North American Indians and Eskimos. *Proceedings of the American Philosophical Society* 96.4: 452–463.

Swadesh, Morris. 1955. Towards greater accuracy in lexicostatistic dating. *International Journal of American Linguistics* 21.2: 121–137.

Trask, Robert L. 2000. *The Dictionary of Historical and Comparative Linguistics*. Edinburgh: Edinburgh University Press.

de Vaan, Michiel. 2008. *Etymological Dictionary of Latin and the Other Italic Languages*. Leiden and Boston: Brill.

Wáng Hóngjūn 王洪君. 2009. *Jiāngù yǎnbiàn, tuīpíng hé céncì de Hànyǔ fāngyán lìshǐ guānxì móxíng* 兼顾演变、 推平和层次的汉语方言历史关系模型 [A historical relation model of Chinese dialects with multiple perspectives of evolution, level and stratum]. *Fāngyán* 方言 3: 204–218.

Wurm, Stephen A. and Liú Yǒngquán 刘涌泉 (eds.). 1987. *Zhōngguó yǔyán dìtújí* 中国语言地图集 [Language atlas of China]. Hongkong: Longman Group.