Improving Phylogeny-Based Network Approaches to Investigate the History of the Chinese Dialects[1]

Johann-Mattis List

Phylogeny-based network approaches are a powerful tool to study language history. Based on a reference tree, they infer the minimal amount of transfer events that is needed to explain the patterning of cognate sets observed in contemporary languages. Since these approaches handle both vertical and lateral aspects of language history, they seem perfectly adequate to study Chinese dialect history. In this paper, a couple of modifications to previous phylogeny-based network approaches are presented. Having confirmed that these modifications constitute significant improvements by testing them on a control-dataset of 40 Indo-European languages, the new method is applied to a dataset of 40 Chinese dialects. The results show that the majority (60%) of character patterns in the Chinese dataset cannot be readily explained as resulting from vertical inheritance alone, much more than observed for the Indo-European data (32%). Since the method yields concrete assessments regarding the regularity of cognate sets, it is very useful as a starting point for deeper historical analyses.

1 Trees, Waves, Networks, and Chinese Dialects

The sociolinguistic situation in China is unique and the history of the various linguistic varieties spoken in China is incredibly complex. It is not surprising that many scholars claim that the family tree model (Schleicher 1853) is inadequate to model Chinese dialect history (Norman 2003, Sagart 2001), since it ignores the horizontal dimension of language relations that played such an important role for the development of the dialects into their current shape. Unfortunately, the alternative model, the *Wave theory* (Schmidt 1872), is also not very helpful, since it ignores the vertical dimension of language relations that is – of course – also constituent for the history of the Chinese dialects. Network models show a way out of the dilemma, since they can be easily used to display both vertical *and* horizontal language relations, as illustrated early by Southworth (1964) for the Indo-European languages, and in a recent paper by Wáng (2009) for the Chinese dialects. The resulting networks are often called *phylogenetic networks*, but following Morrison (2011: 42), I prefer to call them *evolutionary networks*, since these networks claim to display direct hypotheses regarding the phylogeny of the taxonomic units they represent. For this claim to be possible, evolutionary networks need to have a *root* and *internal nodes* that represent ancestral states of the taxonomic units (such as proto-languages in linguistic applications).

Phylogeny-based network approaches (List et al. forthcoming, Nelson-Sathi et al. 2011) are automatic approaches to network reconstruction that come quite close to true evolutionary networks, since they handle both vertical and horizontal language relations. Given a reference tree and a set of words clustered into cognate sets, these methods yield concrete historical scenarios and predict which of the cognate sets has probably been affected by borrowing during its history. Since the methods yield concrete scenarios, their results can be directly checked or used as basis

[1] This study was supported by by the ERC starting grant 240816 “Quantitative modeling of historical-comparative linguistics”. I am very grateful to Prof. Laurent Sagart who not only provided the reference tree that was used in this study, but also many helpful comments.

for deeper research. In the following, I will present how these approaches can be further improved, and how their application to Chinese dialect data can serve as a starting point to investigating Chinese dialect history.

2 Reconstruction of Phylogenetic Networks

2.1 *Distance- and Character-Based Approaches*

It is common to distinguish between distance- and character-based methods for phylogenetic reconstruction. The main difference between these different families of methods lies in the aggregation of information: distance-based methods aggregate information on the taxonomic level. Similarities and differences between all taxonomic units (language varieties) are reduced to distance scores. Character-based methods aggregate information on the level of the items that are selected to define the taxonomic units. Character-based methods yield concrete, individual evolutionary scenarios for each character in the dataset.

The most popular distance-based methods for phylogenetic network reconstruction are based on the technique of split decomposition (Huson et al. 2010: 87-126) as implemented within the SplitsTree software package (Huson 1998). These methods are quite popular in historical linguistics and have been used in a lot of studies on different language families (Bryant et al. 2005, Hamed 2005, Hamed and Wang 2006). However the new insights these methods provide are rather limited. Only very general conclusions regarding the tree-likeness of the data can be drawn and the results are extremely difficult to interpret. Neither can rates of borrowing be calculated, nor can individual borrowing events be inferred.

Character-based methods for phylogenetic network reconstruction are still in their infancy. In a pilot study by Nelson-Sathi et al. (2011) a phylogeny-based method that was originally designed to study microbial evolution (Dagan et al. 2008) was used to assess borrowing frequencies during Indo-European language history. In List et al. (forthcoming), an improved version of this approach was applied to Chinese dialect data. In contrast to distance-based approaches the new approaches infer concrete evolutionary scenarios for all characters in a dataset. The results of the analysis can be easily visualized by combining a reference tree reflecting vertical inheritance with the lateral connections inferred by the method. In contrast to early linguistic proposals to combine the tree and the wave model of language evolution in network models (Southworth 1964) the phylogenetic networks reconstructed by this approach are substantiated both formally and quantitatively.

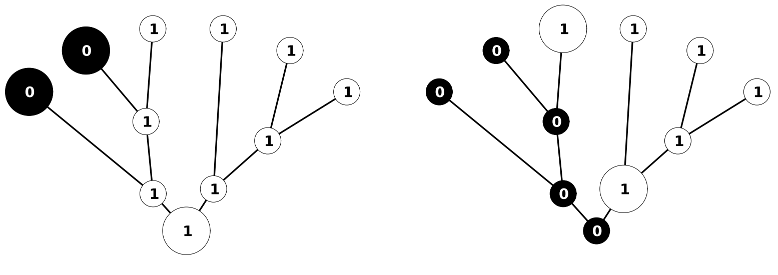
2.2 *Phylogeny-Based Reconstruction of Phylogenetic Networks*

The phylogeny-based method employed in Nelson-Sathi et al. (2011) and List et al. (forthcoming) takes as input a *reference tree* and a set of *phyletic patterns*. Phylogenetic networks are inferred within a three-stage approach. In the first stage, *gain-loss mapping techniques* are used to infer a range of different *gain-loss models* that explain how the cognate sets could have. In a second stage, the best model is chosen by comparing the ancestral and the contemporary *vocabulary size distributions*. In the third stage, a *minimal lateral network* is reconstructed from the gain-loss scenarios inferred by the best model.

Gain-loss mapping (GLM, Cohen et al. 2010, Mirkin et al. 2003) is a standard technique in evolutionary biology. It is used to test the tree-likeness of a given dataset and to infer lateral gene transfer events. The basic goal of all GLM approaches is to infer *gain-loss scenarios* that explain how a given phyletic pattern developed along a reference tree. A phyletic pattern is a matrix representation of the distribution of cognate sets in a given set of languages. The matrix displays whether cognate sets have reflexes in a given language or not. For each language a given cognate set is represented by two states: *presence* (1) or *absence* (0). Depending on the cognate sets being investigated, different patterns can be observed. This is illustrated in Table 1 where translations of "to count" in three Romance and three Germanic languages are split into two cognate sets and coded as phyletic patterns. Given a reference tree that reflects the general evolution of the languages, a *gain-loss scenario* (GLS) explains the evolution of a character in terms of *events* (state changes from ancestral to descendant nodes of the reference tree), with *gain events* being defined as changes from state 0 to state 1, and *loss events* being defined as changes from state 1 to state 0. Figure 1 shows two possible gain-loss scenarios for the first phyletic pattern from Table 1. In scenario (a), one gain event and two loss events are inferred. The scenario thus implies that English *count* was inherited from the common ancestor of the Romance and the Germanic languages with its reflexes being lost in German and Danish. Scenario (b), however, implies that no ancestral form of *count* was present in the common ancestor of the Germanic languages and that it originated independently in English and the ancestor of the Romance languages. We know, of course, that the second scenario is the right one, since English *count* was borrowed from Old French *conter*. Since the independent origin of characters in different branches of a language family is rather rare, we can make the (simplifying) assumption that *patchy cognate sets*, i.e. cognate sets for which a given gain-loss scenario suggests multiple gain events, are the result of language contact.

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

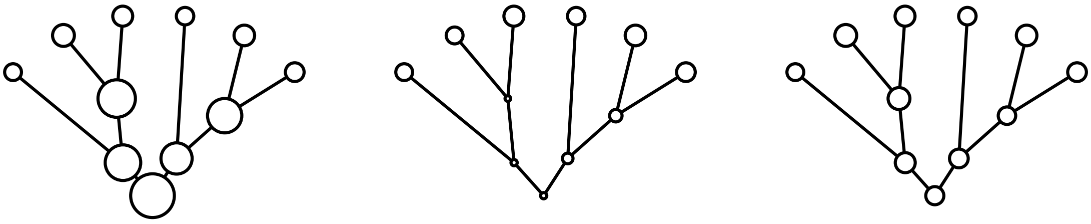
Table 1: Phyletic patterns for "to count" in Romance and Germanic languages.



(a) one gain, two losses (b) two gains, no losses

Figure 1: Gain-loss mapping analyses of reflexes of Latin *computare*.

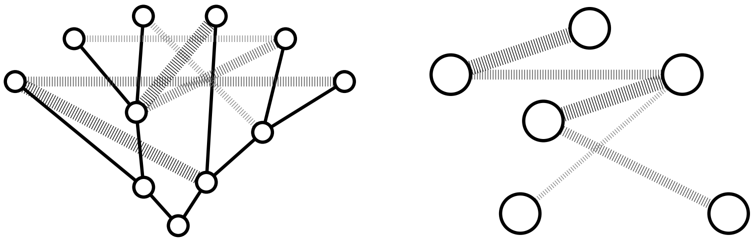
In order to find an appropriate gain-loss scenario for a given phyletic pattern, it is important to find criteria that define the appropriateness of gain-loss scenarios. Only *knowing* that of the two scenarios in Figure 1 the second is the right one does not help us to select the correct scenario in cases where we don't know the history of the languages in such great detail. Internally, it is easy to define *gain-loss models* that favor one of the scenarios by either restricting the maximal number of gain events (*restriction-based approaches*, Nelson-Sathi et al. 2011), or by defining specific penalties for gain and loss events (*parsimony-based approaches*, List et al. forthcoming). Externally, however, specific criteria are needed to determine the best model for a given dataset. Nelson-Sathi et al. (2011) follow Dagan and Martin (2007) in using *ancestral vocabulary size distributions* as a heuristic to determine an optimal gain-loss model. The vocabulary size distribution (VSD) of a given language is defined as the number of words the language uses to express a given set of concepts. The basic idea of the approach is that the number of words that are used to express a given number of concepts in ancestral languages should not differ greatly from the number of words used to express the same concepts in contemporary ones. As illustrated in Figure 2, models that overestimate the tree-likeness of the data yield ancestral VSDs that grow drastically (a), while models that overestimate the amount of lateral transfer yield drastically shrinking VSDs (b). The preference should be given to models that yield well-balanced VSDs throughout all nodes of the tree (c).



(a) growing VSD (b) shrinking VSD (c) well-balanced VSD

Figure 2: Vocabulary size distributions for different gain-loss models (see text).

Having determined an appropriate gain-loss model for a given dataset, the results of the GLM analysis can be interpreted and further analyzed in different ways. The simplest way is to sort out all patchy cognate sets and to investigate these cases individually. Patchy cognate sets can have different origins: They may result from (a) independent convergent evolution, (b) any form of contact between the involved languages or their descendants, including direct, but also semantic transfer, or (c) errors in the data. Given that independent convergent evolution is not a very frequent process (neither in biology nor in linguistics) and that errors in the data should not occur in an ideal world, it is straightforward to assume that the patchiness of the cognate sets results from contact. For a global representation of all patchy cognate sets inferred for a given dataset, one can reconstruct a *minimal lateral network* (MLN, Dagan et al. 2008, Nelson-Sathi et al. 2011). An MLN displays patterns of vertical and lateral inheritance. The reference tree represents vertical relations. Additional edges drawn between the nodes represent the number of times multiple gain events were inferred (see Figure 3a). A specific case of a minimal lateral network, the *minimal spatial network* (MSN) was introduced in List et al. (forthcoming). An MSN represents the lateral edges between the contemporary languages in geographical space with links inferred between ancestral nodes being attributed to the geographically closest descendants (see Figure 3b).



(a) MLN (b) MSN

Figure 3: Minimal lateral networks and minimal spatial networks.

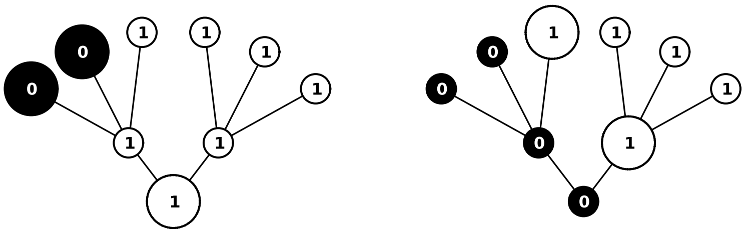
3 Improving Phylogeny-Based Network Reconstruction

Although the original method by Nelson-Sathi et al. (2011) works quite well in assessing the general tree-likeness of a dataset, it has a couple of obvious shortcomings. The gain-loss mapping approach used by the method is a rather simple top-down algorithm that restricts the number of gain-loss models which can be tested and also leads to an overestimation regarding the patchiness of the data. The modifications proposed in List et al. (forthcoming) cope with this by employing a parsimony-based bottom-up approach that makes it possible to reconstruct more fine-graded gain-loss scenarios. Nevertheless, there is still a lot of space for improvement. In the following, I will introduce a couple of modifications to the original approaches that increase both the applicability and the realism of phylogeny-based approaches to phylogenetic network reconstruction. All modifications are implemented as part of LingPy (version 2.1.dev), a Python library for quantitative tasks in historical linguistics (List and Moran forthcoming). LingPy does not not only offer basic algorithms for the tasks described in this paper, but also novel routines to visualize the results. All plots in this paper were done with help of the library.

3.1 *Multifurcating Reference Trees*

So far, all linguistic approaches to gain-loss mapping (List et al. forthcoming, Nelson-Sathi et al. 2011) require *bifurcating* reference trees as input. The requirement for bifurcating trees is a less pending problem in biological applications where the datasets are much larger and bifurcating phylogenies are usually reconstructed automatically. In linguistics, however, where scholars are very cautious when it comes to proposing detailed phylogenies, multifurcating reference trees are the rule rather than the exception, and phylogeny-based network reconstruction methods should definitely be able to handle them.

Multifurcation does not constitute a theoretical problem for the idea of gain-loss mapping. As can be seen from the examples in Figure 4, it makes no huge difference whether characters are mapped on a bifurcating or a multifurcating reference tree. Algorithmically, however, handling multifurcation can be quite challenging and, depending on the underlying algorithm, even impossible. The restriction-based top-down approach by Nelson-Sathi et al. (2011), for example, cannot be extended to multifurcating reference trees, since the GLM method is theoretically stated as a binary search procedure (Dagan and Martin 2007). The parsimony-based bottom-up approach by List et al. (forthcoming), however, employs an exhaustive search for all possible gain-loss scenarios. Adjusting it for multifurcation therefore only requires to adjust the search procedure. This was done in the new version of the phylogeny-based network reconstruction method presented in this paper, and multifurcation is now supported as a default.[2]



(a) one gain, two losses (b) two gains, no losses

Figure 4: Gain-loss mapping analysis for multifurcating reference trees.

3.2 *Mixed Gain-Loss Models*

As mentioned above, the GLM method by Nelson-Sathi et al. (2011) employs a restriction-based top-down algorithm. The drawback of this approach is that the number of origins allowed by a given model remains fixed regardless of the input data. The parsimony-based bottom-up approach List et al. (forthcoming) allows for more flexible models by assigning specific penalties for gain and loss events. The advantage of this approach is that the number of models that can be tested on a given dataset greatly increases and that the number of gain events is no longer restricted. Despite these advantages, both approaches still have the drawback that all phyletic patterns in a given dataset are explained with help of one and the same gain-loss model. Given that words differ regarding their stability and borrowability, depending on the concepts they denote, it seems to make much more sense to select gain-loss models on the basis of *single concepts* rather than *overall tendencies* in the data. As a result, the analysis of a given dataset may contain a mix of different gain-loss models, depending on the concepts denoted by the cognate sets. Adjusting the gain-loss mapping algorithm to account for mixed models is straighforward: Instead of using the VSD criterion to determine the best gain-loss model for *all cognate sets*, we now use the same criterion to determine the best model for *all cognate sets* that *denote the same concept*.[3]

Apart from being much more flexible than the previous approaches, this procedure has another great advantage, in so far as mixed gain-loss models yield *explicit statements* regarding lexical change processes, since they reunify the cognate sets under semantic categories that were formerly split by the binary coding procedure. This is illustrated in Figure 5 where gain-loss scenarios inferred for the two phyletic patterns from Table 1 are combined within an explicit framework that shows how the words for the concept “to count” evolved in our small sample of three Germanic and three Romance languages. Note that the evidence from the contemporary languages does

[2] It would go beyond the scope of this paper to go into the details of implementation here. The interested reader is therefore referred to the API and the source code of the LingPy library available from `http://lingpy.org`.

[3] For the details of the evaluation procedure, see our description in List et al. forthcoming.

not allow to reconstruct the state of the ancestor of both language families. For this reason, no character is assigned to the root node.

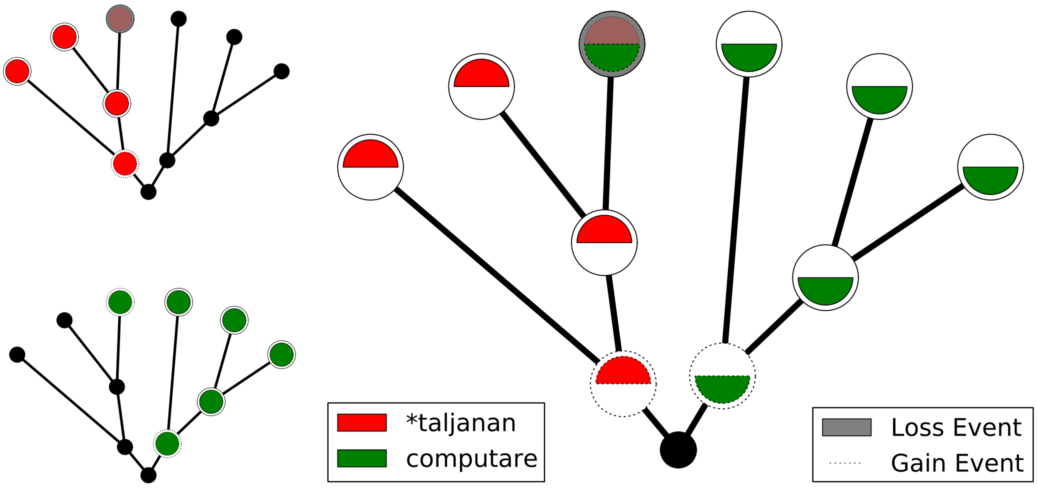


Figure 5: Using mixed models to combine individual gain-loss scenarios.

3.3 *Further Modifications*

Two further smaller modifications were added to the new framework of phylogeny-based network reconstruction. The first one handles the selection of gain-loss scenarios that yield identical scores according to the parsimony-based approach. While in the former approach in these cases the preference was given to scenarios that involved a *minimal* amount of gain events, it is now given to those scenarios involving a *maximal* amount. As a result, gains are *pushed* to the leaves of the tree instead of being pushed to the root. Pushing gains to the leaves has the advantage of moving them to nodes that are easier to observe for the researcher. While, for example, the proposal of archaic borrowing relations between Proto-Germanic and Proto-Romance comes close to speculation, albeit a sophisticated one, recent borrowing relations are far more easy to observe and to verify. Note that the procedure does *not* change the overall amount of borrowings inferred by the method, since the selection of gain-loss models is handled independently.

As a second modification, the new framework now allows for *multiple gain events* in a lineage. In List et al. (forthcoming), the possibility of characters to be gained, lost, and gained again was deliberately excluded in order to restrict the search space. Given that loss and regain via lexical transfer is not impossible,[4] the new framework now allows the user to specify a maximal amount of multiple gains in a lineage, with the default being set to 2.

[4] Compare English *flower*, which was borrowed from Old French *flour*, replacing its cognate form Old English *blostm*.

4 Testing the New Approach

In order to find out whether the modifications proposed in this paper are real improvements to phylogeny-based network reconstruction, it is important to test them. For this purpose, a subset of 40 Indo-European languages was taken from the *Indo-European Lexical Cognacy Database* (IELex, Dunn 2012). The subset contains 207 glosses corresponding to 7 518 words clustered into 1 194 cognate sets (see Supplemental Material I.5). Of the originally 9 413 words, 1 895 were excluded as *singletons*, since they could not be shown to be cognate with any other word in the sample. The advantage of this dataset is that known borrowings are marked along with their sources. This gives us the possibility to treat the known borrowings as cognates and to test whether the method correctly identifies these cognate sets as patchy cognates or not. The data was further modified by correcting for obvious errors in the cognate judgments and introducing more, so far unobserved known cases of borrowing, especially in the Slavic and the Romance branch of the languages in our sample. This yielded a total of 186 known cases of borrowing that are cognate with at least one of the other words in the dataset, and a total of 100 cognate sets in which at least one of the 186 words occurs (see Supplemental Material I.6). Two phylogeny-based network reconstruction methods were tested, the parsimony-based approach by List et al. (forthcoming), and the modified approach with mixed gain-loss models presented in this paper. Both methods were tested on two different reference trees, one bifurcating one, with the major subgroupings following the analysis of Ringe et al. (2002), and one multifurcating one, following the major subgroupings of Ethnologue (Lewis and Fennig 2009) with the exception that Slavic and Baltic were assigned to a common branch. For the parsimony-based approach, nine different models were tested, with gain-loss ratios ranging between 1:1 and 3:1 in steps of 0.25 (1:1, 5:4, 3:2, ..., 11:4, 3:1),[5] and the model that yielded the highest p-value in the Wilcoxon rank-sum test (Kruskal 1957) of contemporary and ancestrals VSDs was selected as the best one. The nine models were also taken as basic models for the mixed GLM approach, but in contrast to the old analysis, the modified version of the algorithm that allows for two gains in a lineage and pushes gains to the leaves was used.

Reference Tree	**bifurcating**		**multifurcating**	
Method	**OLD**	**NEW**	**OLD**	**NEW**
(Best) Gain-Loss Model	5:2	mixed	2:1	mixed
Overall p-Value	0.98	0.54	0.86	0.98
Number of Origins	1.30	1.36	1.40	1.43
Proportion of Patchy Cognates	0.25	0.29	0.30	0.32
Correctly Identified Patchy Cognates	0.59	0.68	0.71	0.77

Table 2: Comparing the new approach with the old one.

The results of this analysis are displayed in Table 2 (see also Supplemental Material I.7). As can be seen, both the application of mixed models and the use of multifurcating reference trees enhance the results greatly, with 18% differences between the old approach applied to the bifurcating reference tree and the new approach applied to the multifurcating one. These differences are significant with $p < 0.01$, using the Wilcoxon singed rank test. Furthermore, of the 23 cases

[5] See List et al. forthcoming for details regarding the definition gain-loss models.

where the method fails to detect a patchy cognate set, at least 8 cases are very hard (if not impossible) to detect by the phylogeny-based approach, since they are regularly reflected in most of the taxa, involving only two loss events, such as English *die*, being reflected in Frisian and all Scandinavian languages, but missing in German and Dutch. The resulting MLN is displayed in Figure 6. The inferred connections reflect known contact relations between the Indo-European languages quite well. Thus, one of the heaviest edges in the MLN connects Albanian, which borrowed many words from Latin (Orel 2000: 23f), with the ancestor of all Romance languages. English, which was heavily influenced during its history by both Scandinavian and Norman French (Harbert 2007: 23f), has heavy links with the ancestor of Scandinavian and the ancestor of Romance. And Armenian, which is assumed to have been in close contact to Greek in prehistorical times (Schmitt 1981[2007]: 22f), shares a heavy link with Greek. Even if it does not identify all cognate sets with known borrowings in our testset, the method is quite good at detecting major tendencies.

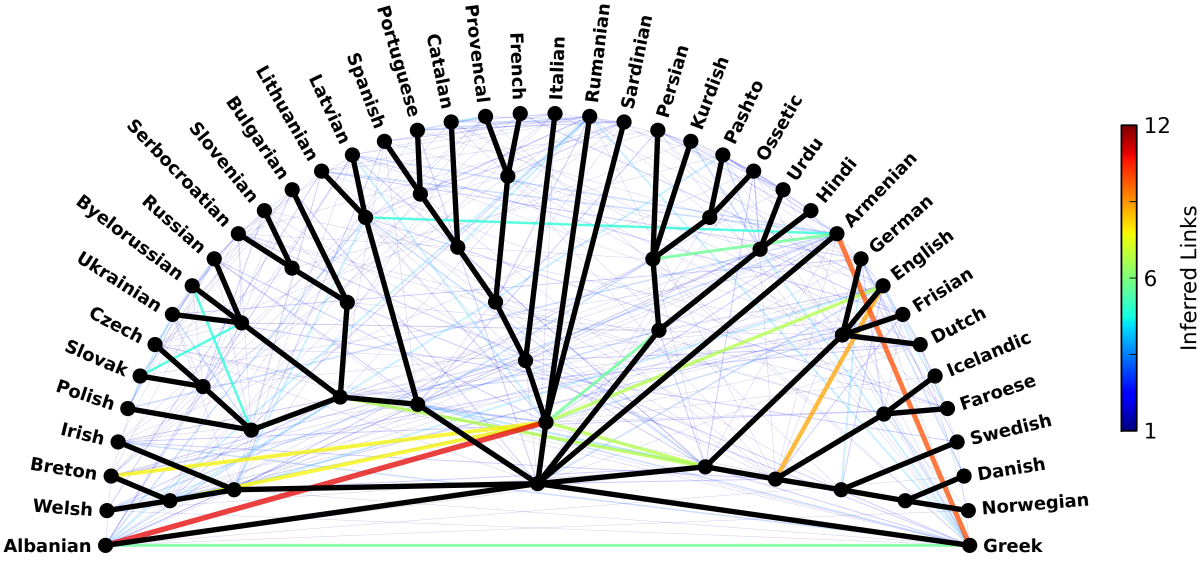


Figure 6: The minimal lateral network of the Indo-European data (mixed GLM).

5 Application to Chinese Dialect Data

5.1 *Materials and Methods*

The data used in this test was taken from a cleaned digital version (List et al. forthcoming) of the *Xiàndài Hànyǔ fāngyán yīnkù* (Hóu 2004), consisting of 180 concepts translated into 40 Chinese dialect varieties. In contrast to the version used in List et al. (ibid.), some errors resulting from the semi-automatic cleaning procedure could be found and corrected. As a result, the current version of the dataset consists of a total of 7 952 words clustered into 1056 cognate sets. Of the originally 9 957 words, 2005 were excluded, since they could not be found to be cognate with any other word in the sample. All the data, including a list of the taxa, the glosses, and the cognate assignments, is given in Supplemental Material II.

A major problem for the application of the phylogeny-based network reconstruction approach to Chinese dialect data is that it requires a reference tree as input. Due to the complex linguistic situation in China, the history of the major dialect groups is still much disputed, and none of the many subgroupings that have been proposed so far (cf., e.g., Karlgren 1954, Lǐ 2005, Norman 2003, Wáng 2009) has gained general acceptance. For this test, I decided to take a recent work-in-progress proposal by Laurent Sagart (personal communication) which has the advantage that it is explicitly historically oriented, being substantiated by distinct innovations for each split in the family tree. Since Sagart's proposal so far only includes the seven major dialect groups, leaving the three transitional groups of Jìn, Pínghuà, and Huī unassigned, I followed his innovations and reassigned the three groups accordingly. The family tree for the ten major dialect groups is given in Figure 7. For the internal classification of the major dialect groups, I generally followed the groupings proposed in the *Language Atlas of China* (Wurm and Liu 1987). However, in certain cases, where these groupings were too shallow and additional information was available, the internal subgrouping was further modified. Thus, 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 with help of quantitative analyses of the data and suggestions from the literature (Ting 1991). The full reference tree, including a list of the innovations proposed by Sagart, is given in Supplemental Material II.4.

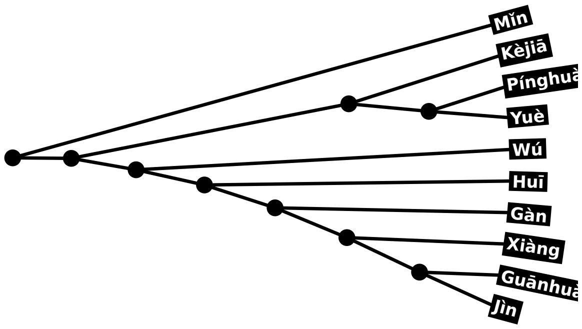


Figure 7: Reference tree of the major dialect groups based on Sagart's classification.

The data was analyzed with help of the improved approach to phylogeny-based network reconstruction, using the same settings that were also used for the analysis of the Indo-European data described in Section 4. Thus, nine different gain-loss models were taken as basis for the mixed GLM approach. In addition to the *minimal lateral network*, a *minimal spatial network* was also reconstructed, following the description in List et al. (forthcoming).

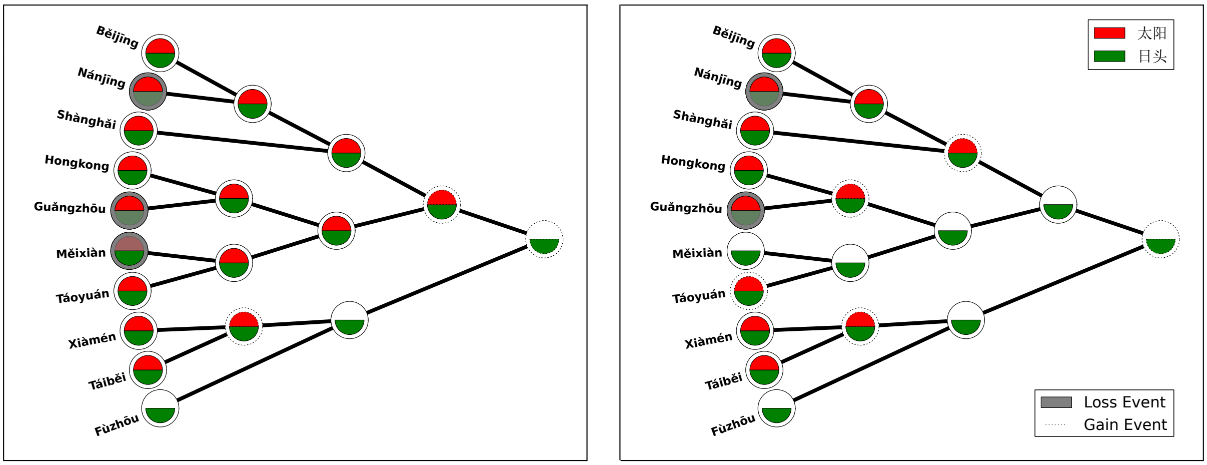
5.2 *Results*

Table 3 lists some general statistics for all gain-loss models that were tested. As can be seen from the table, the mixed model achieves the highest overall p-value, yielding a vocabulary size distribution of the ancestral varieties that comes closest to the vocabulary size distribution of all contemporary varieties in the sample. The difference between the distributions of origins per cognate proposed by the mixed model and the non-mixed model with the highest p-value (9:4) is significant with $p < 0.01$, using the Wilcoxon signed rank test. According to the mixed model, 61% of all cognates in the data are patchy, with the average number of origins being 2.01. Comparing these scores with those inferred for the Indo-European data (32% of patchy cognates and 1.43 origins on average, see Table 2), the differences are quite striking. However, one should keep in mind that the datasets are only partially comparable. In contrast to the Indo-European dataset, only a small proportion of glosses in the Chinese dataset belongs to the realm of basic vocabulary. When splitting the Chinese dataset into a "basic" and a "non-basic" part, consisting of 48 and 132 glosses, respectively, there are 1.89 origins on average in the basic part, and 2.05 origins in the non-basic part. The differences, however, are not significant with $p = 0.16$ (using the Wilkoxon rank sum test), and the proportions of patchy cognates differ only slightly (60% vs. 61%).

Gain-Loss Model	**3:1**	**11:4**	**5:2**	**9:4**	**mixed**	**2:1**	**7:4**	**3:2**	**5:4**	**1:1**
Number of Origins	1.74	1.74	1.80	1.86	2.01	2.20	2.25	2.39	2.47	3.20
Maximal Number of Origins	7	7	7	9	9	9	10	13	13	14
Overall *p*-Value	0.08	0.08	0.23	0.50	0.78	0.11	0.04	< 0.00	< 0.00	< 0.00
Proportion of Patchy Cognates	0.52	0.52	0.54	0.55	0.61	0.66	0.66	0.68	0.68	0.11

Table 3: Comparing the gain-loss models.

That the application of mixed gain-loss models results in actual improvements of the analysis is shown in Figure 8, where parts of the scenarios for *tàiyáng* 太阳 and *rìtóu* 日头"sun" as inferred by the 9:4 model and the mixed model are displayed. While the 9:4 model infers only two distinct origins for *tàiyáng* 太阳, one in the Mǐn, and one in the ancestor of all non-Mǐn dialects, the mixed model infers four separate origins in Mǐn, Hakka, Yuè, and the common ancestor of Wú and Mandarin. Despite the fact that both models fail to detect that *tàiyáng* 太阳 is a very recent borrowing from Standard Chinese, be it in Wú, Yuè, Hakka, or Mǐn, the mixed model comes much closer to the truth than the non-mixed model.



(a) general model (9:4)

(b) mixed model

Figure 8: Comparing general models with mixed models.

The minimal lateral network of the analysis is shown in Figure 9. In contrast to the MLN of the Indo-European languages, where as many as eight of the ten heaviest lateral links involved at least one internal node (see Supplemental Material I.9), we find a somewhat opposite situation here, with seven of the ten heaviest links being drawn between two external nodes. Six of these links occur between Northern Mandarin (Lányín and Zhōngyúan) and Jìn, three links involve Southern Mandarin (Xīnán and Jiānghuái), and only one heavy link of seven cognates between Hǎikǒu and the ancestor of all non-Mǐn dialects involves a dialect group other than Mandarin or Jìn (see Supplemental Material III.5). The reference tree classifies the Jìn dialects as the first outgroup of the Mandarin branch. Given that the general position of the Jìn dialects is highly disputed with quite a few scholars favoring a more internal position (Ting 1991, Yan 2006), it is hard to say whether the links inferred between Jìn and Mandarin result from contact or genetic closeness greater than suggested by the reference tree. When moving the Jìn dialects on the reference tree to an internal position where they are grouped as direct siblings of Lányín and Zhōngyuán Mandarin, most of the patchy links disappear, but the overall proportion of patchy cognates decreases only slightly from 61% to 60% with 2.02 origins on average (see also Supplemental Material III.6). However, reference trees that decrease the amount of patchiness for a given language or language group do not necessarily reflect historical reality. When treating English as an outgroup of all Germanic languages, for example, its close contact-induced connection to the Scandinavian languages is no longer detectable by phylogeny-based network reconstruction methods, since true historical processes are masked by the reference tree (Nelson-Sathi et al. 2011). Therefore, we cannot solve the question here. We have to wait until further research sheds more light on the historical relations between Mandarin and Jìn.

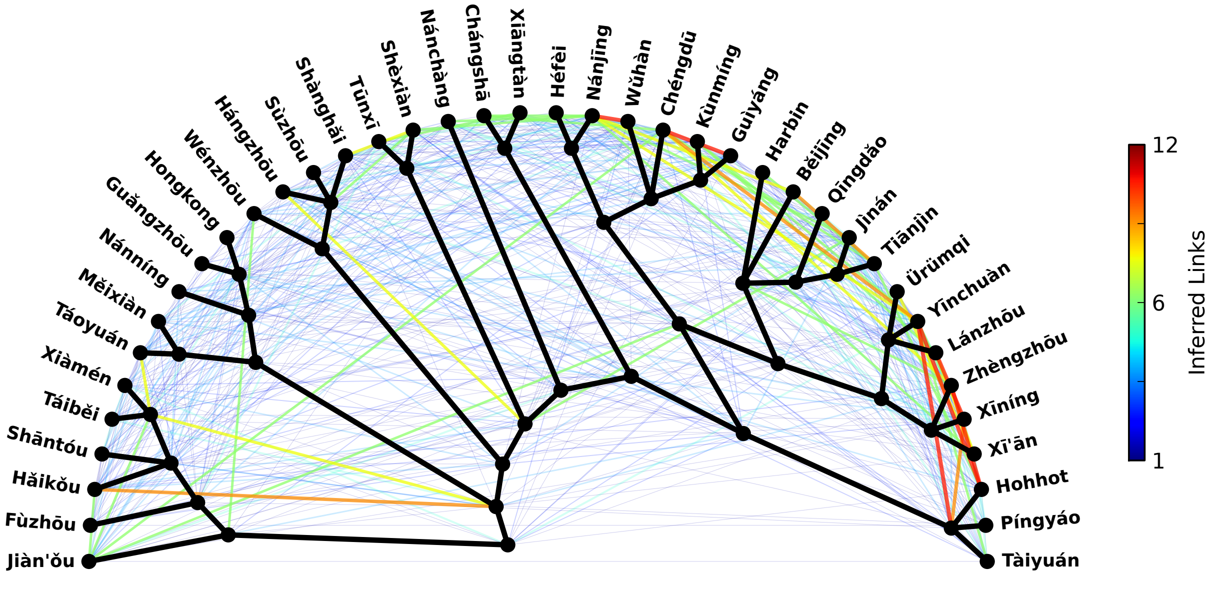


Figure 9: The minimal lateral network of the Chinese dataset.

There are five heavy edges (with ≥ 6 cognates) that involve what Norman (2003) called the Central (Wú, Huī) and the Southern dialects (Hakka, and Mǐn). These edges are listed in Table 4 along with the patchy cognates that constitute them. It is not clear to which degree the patchiness of the words in the table is due to real borrowing processes or other reasons, such as independent semantic shift, or even errors in the coding. Nevertheless, the majority of glosses listed in the table belong to a class of concepts that are generally easy to borrow, such as food names ("baozi"), vegetable names ("spinach", "cabbage"), and artifacts in a broad sense ("spoon"). Furthermore, none of the connections inferred by the method is specifically surprising. Hǎikǒu, for example, is a Mǐn dialect that is geographically isolated from the other Mǐn dialects and close to many other dialect groups, especially Yuè and Pínghuà, but also Mandarin and Hakka. The patchy cognates listed in Table 4 are all rather untypical for a Mǐn dialect which suggests that these words were borrowed into the Hǎikǒu lexicon. Táoyúan is a variety of Hakka spoken on the Taiwan island. Due to the predominance of Mǐn dialect varieties (such as Táiběi) in this region, the Hakka varieties have been heavily influenced by these (Lín 2012). Of the six inferred patchy cognates, 豆油 "soya sauce" and 对 "from", are only reflected in Táiběi and Xiàmén. This suggests that the words were borrowed from Táiběi into Táoyuán. Hángzhōu was classified as a Wú dialect in the reference tree, following the traditional classification of the *Language Atlas of China* (Wurm and Liu 1987). However, this classification has been challenged by alternative proposals according to which Hángzhōu is more likely to be a geographically displaced Mandarin dialect (Simmons 1995). No matter which of the theories is right, Hángzhōu's closeness to the Northern dialects finds its reflection not only in the heavy link given in the table, but also in a large number of less heavy links (3 to 4 cognates), connecting it either directly to the Mandarin subgroup, or its earlier ancestors (see Supplemental Material III.5).

Nodes		**Weight**	**Cognates**
Hǎikǒu	non-Mǐn	7	刚刚 "just (just came)", 淡 "light", 南瓜 "pumpkin", 菠菜 "spinach", 勺 "spoon", 瘦 "thin", 从 "from"
Táiběi, Xiàmén	non-Mǐn	6	只 "only", 中秋节 "Mid-Autumn Festival", 房间 "flat", 只 classifier (cow), 冷 "cold", 只 classifier (pig)
Táiběi, 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"

Table 4: Heavy links between Central and Southern dialects.

Although minimal lateral networks are a clear improvement over distance-based networks, they should not be confused with true *evolutionary networks* in the sense of Morrison (2011: 42), since they do not display *direct* phylogenetic hypotheses. Minimal lateral networks cannot infer concrete borrowing events, since they can neither detect their direction, nor their source, and it requires some effort to interpret them. Minimal spatial networks, which were first introduced in List et al. (forthcoming), can ease these efforts by presenting minimal lateral networks from a different viewpoint. In these networks, connections between internal and external nodes are drawn between the geographically closest contemporary varieties. As can be seen from the MSN shown in Figure 10, the heavy links given in Table 4 are mostly confirmed. Furthermore, some additional links that were much harder to spot on the MLN become now apparent, such as the connections between Wénzhōu, a Wú dialect, and Jiàn'ǒu and Fùzhōu, two representatives of Mǐn, which are very typical for the Wú-Mǐn dialect border (Pan 1991: 249f).

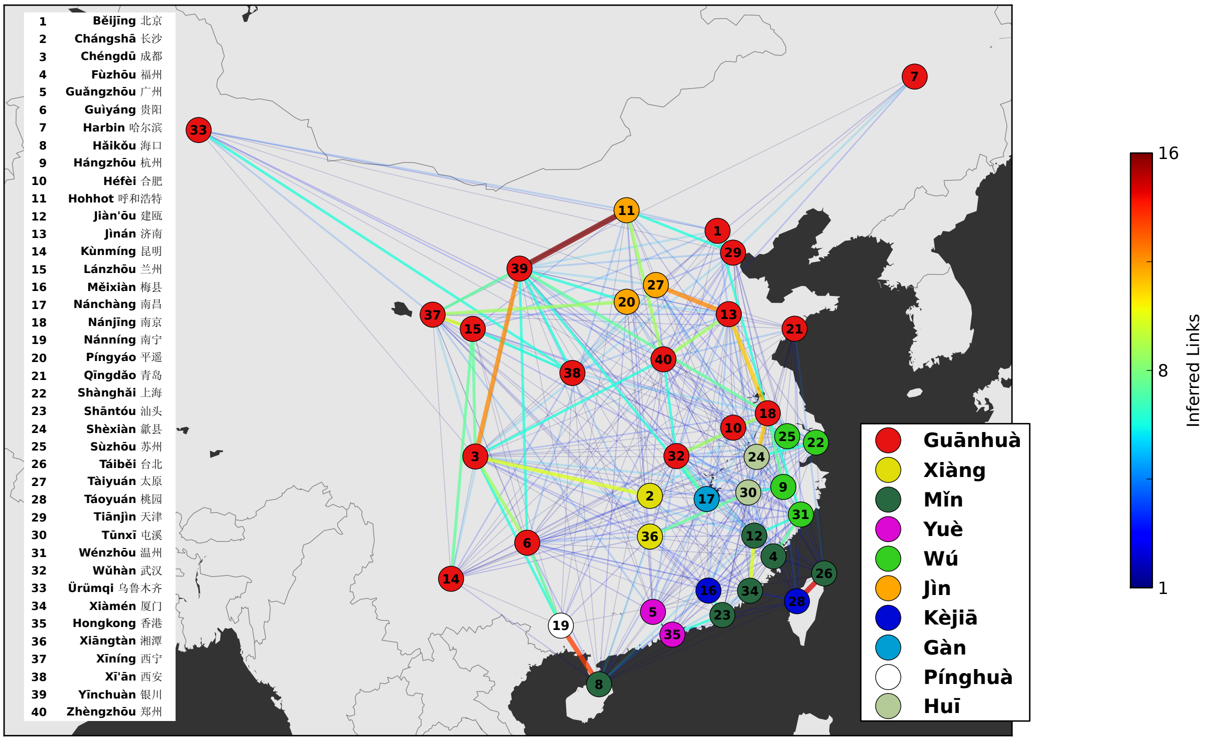


Figure 10: The minimal spatial network of the Chinese dataset.

6 Discussion

Phylogeny-based network approaches offer new possibilities for quantitative historical linguistics. In contrast to traditional methods for phylogenetic tree reconstruction, they handle both vertical and horizontal aspects of language history. In contrast to distance-based approaches to phylogenetic network reconstruction, they yield concrete evolutionary scenarios. In this paper I presented a couple of modifications to these approaches. Testing them on a control-dataset of 40 Indo-European languages showed that they constitute significant improvements over earlier approaches. Using a dataset of 40 Chinese dialect varieties, I further illustrated how the new methods can be employed to investigate Chinese dialect history. Although the approach is not (yet) capable of producing concrete evolutionary hypotheses, since the inferred connections are undirected, it provides valuable assessments regarding the regularity of cognate sets and can therefore serve as a good starting point for deeper historical analyses.

Supplemental Material

The Supplemental Material accompanying this study is divided into three parts. The first part contains the data and the results for the test of the method on Indo-European languages. The second part contains information regarding the reference tree of the Chinese dialects and the data upon which this study was based. The third part contains the results for the test on Chinese dialect data.

References

Bryant, D., F. Filimon, and R. D. Gray (2005). “Untangling our past: Languages, Trees, Splits and Networks”. In: *The evolution of cultural diversity: A phylogenetic approach*. Ed. by R. Mace, C. J. Holden, and S. Shennan. London: UCL Press, 67–84.

Cohen, O., H. Ashkenazy, F. Belinky, D. Huchon, and T. Pupko (2010). “GLOOME: gain loss mapping engine”. In: *Bioinformatics* 26.22, 2914–2915.

Dagan, T. and W. Martin (2007). “Ancestral genome sizes specify the minimum rate of lateral gene transfer during prokaryote evolution”. In: *PNAS* 104.3, 870–875.

Dagan, T., Y. Artzy-Randrup, and W. Martin (2008). “Modular networks and cumulative impact of lateral transfer in prokaryote genome evolution”. In: *PNAS* 105.29, 10039–10044.

Dunn, M., ed. (2012). *Indo-European lexical cognacy database (IELex)*. URL: `http://ielex.mpi.nl/`.

Hamed, M. B. (2005). “Neighbour-Nets Portray the Chinese Dialect Continuum and the Linguistic Legacy of China’s Demic History”. In: *Proc B* 272.1567, 1015–1022.

Hamed, M. B. and F. Wang (2006). “Stuck in the forest: Trees, networks and Chinese dialects”. In: *Diachronica* 23, 29–60.

Harbert, W. (2007). *The Germanic languages*. Cambridge: Cambridge University Press.

Huson, D. H., R. Rupp, and C. Scornavacca (2010). *Phylogenetic networks. Concepts, algorithms, and applications*. Cambridge: Cambridge University Press.

Huson, D. H. (1998). “SplitsTree: analyzing and visualizing evolutionary data”. In: *Bioinformatics* 14, 68–73.

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

Karlgren, B. (1954). “Compendium of phonetics in ancient and archaic Chinese”. In: *Bulletin of the Museum of Far Eastern Antiquities* 26, 211–367.

Kruskal, W. H. (1957). “Historical notes on the Wilcoxon unpaired two-sample test”. English. In: *Journal of the American Statistical Association* 52.279, 356–360.

Lewis, M. P. and C. D. Fennig, eds. (2009). *Ethnologue. Languages of the World*. 17th ed. Dallas: SIL International. URL: `http://www.ethnologue.com`.

List, J.-M., S. Nelson-Sathi, W. Martin, and H. Geisler (forthcoming). “Using phylogenetic networks to model Chinese dialect history”. In: *Language Dynamics and Change*.

List, J.-M. and S. Moran (forthcoming). “An open source toolkit for quantitative historical linguistics”. In: *Proceedings of the ACL 2013 System Demonstrations*. (Sofia, Bulgaria, Aug. 4–9, 2013). Association for Computational Linguistics.

Lín Qīngshū 林清书 (2012). “Táiwān Hànyǔ fāngyán yǔ guóyǔ yǔyán jiēchù yánjiū de zhòngyāo yìyì 台湾汉语方言与国语语言接触研究的重要意义 [Significance in contact study of Chinese dialects in Taiwan with Standard Chinese]”. In: *Journal of Longyuan University* 龙岩学院学报 30.6, 5–25.

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

Mirkin, B. G., T. I. Fenner, M. Y. Galperin, and E. 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”. In: *BMC Evolutionary Biology* 3, 2.

Morrison, D. A. (2011). *An introduction to phylogenetic networks*. Uppsala: RJR Productions.

Nelson-Sathi, S., J.-M. List, H. Geisler, H. Fangerau, R. D. Gray, W. Martin, and T. Dagan (2011). “Networks uncover hidden lexical borrowing in Indo-European language evolution”. In: *Proc B* 278.1713, 1794–1803.

Norman, J. (1991). “The Mǐn dialects in historical perspective”. In: *Journal of Chinese linguistics*: *Languages and Dialects of China*. Ed. by W. S.-Y. Wang, 325–360.

— (2003). “The Chinese dialects. Phonology”. In: *The Sino-Tibetan languages*. Ed. by G. Thurgood and R. LaPolla. London and New York: Routledge, 72–83.

Orel, V. (2000). *A concise historical grammar of the Albanian language. Reconstruction of Proto-Albanian*. Leiden, Boston, and Köln: Brill.

Pan, W. (1991). “An introduction to the Wu dialects”. In: *Journal of Chinese Linguistics. Languages and Dialects of China*. Ed. by W. S.-Y. Wang, 237–293.

Ringe, D., T. Warnow, and A. Taylor (2002). “Indo-European and Computational Cladistics”. In: *Transactions of the Philological Society* 100.1, 59–129.

Sagart, L. (2001). “Vestiges of Archaic Chinese derivational affixes in modern Chinese dialects”. In: *Sinitic grammar. Synchronic and diachronic perspectives*. Ed. by H. Chappell. Oxford: Oxford University Press, 123–142.

Schleicher, A. (1853). “Die ersten Spaltungen des indogermanischen Urvolkes”. In: *Allgemeine Monatsschrift für Wissenschaft und Literatur*, 786–787.

Schmidt, J. (1872). *Die Verwantschaftsverhältnisse der indogermanischen Sprachen*. Weimar: Hermann Böhlau.

Schmitt, R. (1981 [2007]). *Grammatik des Klassisch-Armenischen*. 2nd ed. Insbruck: Insbrucker Beiträge zur Sprachwissenschaft.

Simmons, R. V. (1995). “Distinguishing characteristics of the Hangzhou dialect”. In: *New Asia Academic Bulletin* 11, 383–398.

Southworth, F. C. (1964). “Family-tree diagrams”. In: *Language* 40.4, 557–565.

Ting, P.-H. (1991). “Some theoretical issues in the study of Mandarin dialects”. In: *Journal of Chinese linguistics*: *Languages and Dialects of China*. Ed. by W. S.-Y. Wang, 187–236.

Wurm, S. A. and Y. Liu, eds. (1987). *Zhōngguó yǔyán dìtújí* 中国语言地图集 [Language atlas of China]. Hongkong: Longman Group.

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]”. In: *Fāngyán* 方言 3, 204–218.

Yan, M. M. (2006). *Introduction to Chinese dialectology:* München: LINCOM Europa.