

Neighbour-nets portray the Chinese dialect continuum and the linguistic legacy of China's demic history

Mahé Ben Hamed*

Laboratoire Dynamique du Langage, CNRS UMR 5596, Lyon, France

As with species studied by evolutionary biologists, languages are evolving entities. They can evolve in tree-like patterns, possibly blurred by borrowing, but they can also develop in non-tree-like schemes. For instance, diglossia, as in the case of Chinese, can counterbalance the hierarchical pattern expected from differentiation by internal change associated with isolation by distance of speech communities. Using two lexical datasets, either the basic lexicon supposedly more immune to borrowing or a representative sample of the whole lexicon, we investigate the development pattern of Chinese dialects using a neighbour-net approach, which is an unprejudiced technique for representing object relationships. The resulting graphs are consistent with a dialect continuum shaped by counterbalanced effects of homogenizing diglossia and borrowing versus differentiating spread of speech communities. Historical events and linguistic claims can be mapped on these graphs.

Keywords: Chinese; language change; neighbour-net; dialect continuum; diglossia; borrowing

1. INTRODUCTION

Internal change associated with the isolation of speech communities can be equated with the process of speciation through descent with modification, among isolated populations. The tree metaphor used by Darwin to represent species evolution was inspirational to historical linguistics. The Indo-Europeanist philologist August Schleicher (1853) defined a similar model of language development, the *Stammbaumtheorie* (family-tree theory), which found an early opponent in Schleicher's own disciple (Schmidt 1872). Schmidt argued for the non-hierarchical diffusion of linguistic innovations from multiple sources, for which trees cannot account. More recently, Dixon (1997) argued that, while the tree model may be adequate in the case of the most extensively studied families, such as Indo-European, Semitic, Uralic or Algonquian, there are linguistic innovations that the tree model does not adequately explain. Borrowing is a major type of deviation from the tree-like fashion. In fact, owing to its function, language is very sensitive to contact between speech communities and to social factors, which induce borrowing. Historical linguistics circumvents the problem by selecting linguistic features that are assumed to be more immune to borrowing than the rest of the language. This controlled sampling aims at decreasing the divergence from the tree model and thus at legitimizing its use. However, the question of whether language evolution is tree-like or not is not related to borrowing only. The development of a set of languages can violate the necessary condition of isolation, as in cases of diglossia. Diglossia is defined as a kind of bilingualism between closely, genetically related languages or language varieties (Ferguson 1959). It is associated with a high degree of specialization between

the two levels where each variety is assigned to certain functions. Chinese, for instance, is typical of a diglossic linguistic domain. There are regional variants for which interintelligibility decreases with geographical distance. These spoken variants are supplemented by standard forms that serve for communication across dialect boundaries, bureaucracy and as the written standard for literary and cultural tradition (Yuan 1980). The standard forms, developed from northern dialects, were thus used alongside related, yet historically more evolved, spoken varieties of the same language (Norman 1988). Such a situation would logically create a homogenizing context, counterbalancing tree-like development through isolation by distance.

The history of China has been a succession of migratory waves, populations being driven away from their homelands by war and natural disasters, and immigrants often outnumbering the indigenous populations occupying the lands in which they finally settled (Zhou & You 1986). These massive demic movements have left a rich linguistic legacy. Ogura (1994) defines them as the forces that created the Chinese dialects. In such a context of population dispersion, dialects are expected to differentiate in a tree-like scheme. The differences among Chinese dialects are indeed considerable, and it is often said that they are in fact different languages (Norman 1988). According to Norman, Chinese is more like a language family that has evolved through three millennia, rather than a single language displaying mere regional variations. Chinese is conventionally divided into seven groups (figure 1), following Yuan (1980). These groups' boundaries correspond to the sharpest zones of variation. Mandarin occupies north and western China. From north to south, the coastal dialects are Wu, Min (also spoken in Taiwan and Hainan islands), Hakka (or Kejia) and Yue. The transition between Mandarin and the coastal dialects

*mahe.ben-hamed@ish-lyon.cnrs.fr.

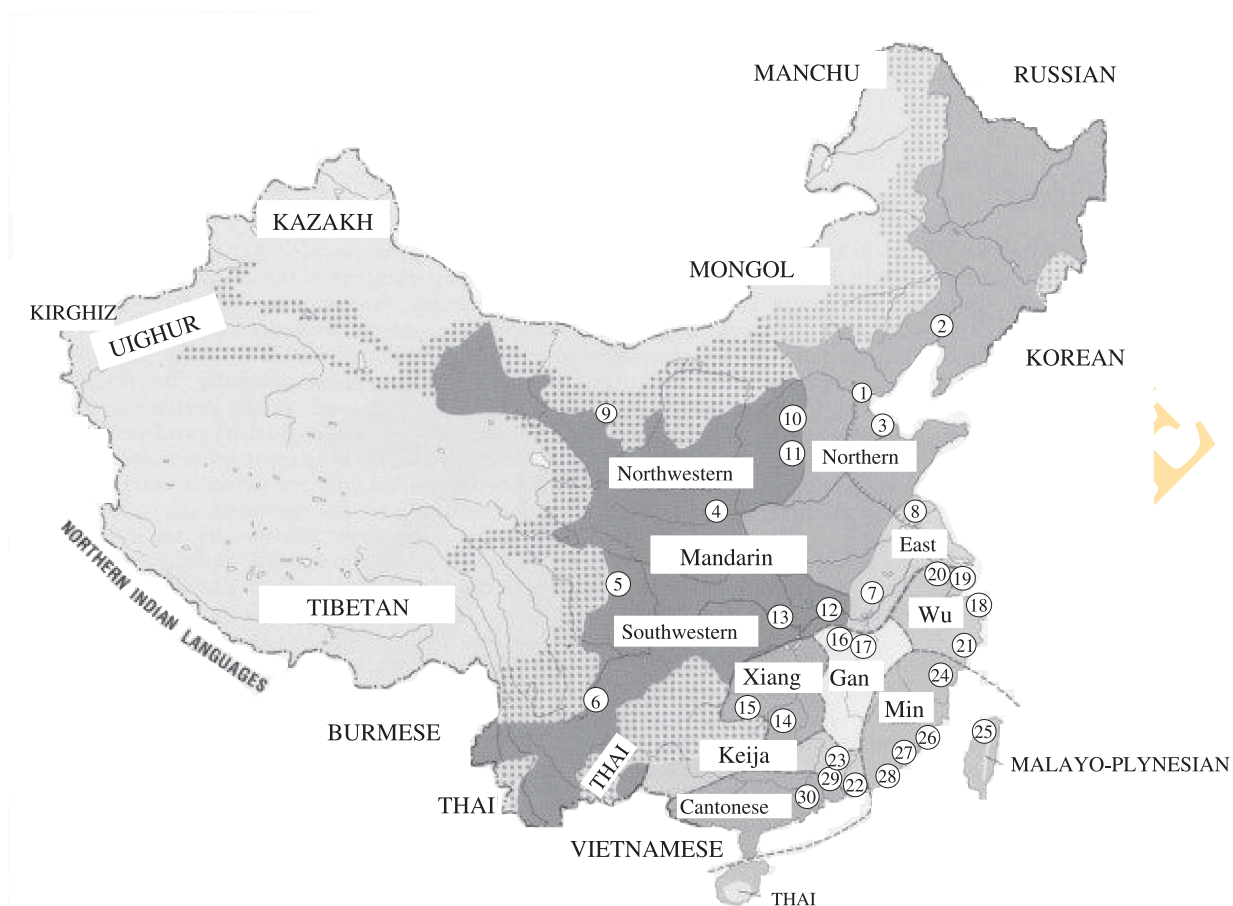


Figure 1. Map of China's dialects, showing the locations in which the dialects used in this study are spoken. *Mandarin*: (1) Beijing, (2) Shenyang, (3) Jinan, (4) Xi'an, (5) Chengdu, (6) Kunming, (7) Hefei, (8) Yangzhou, (9) Ningxia, (10) Taiyuan, (11) Wuhan, (12) Yuci, (13) Yingshan. *Xiang*: (14) Changsha, (15) Shuangfeng. *Gan*: (16) Anyi, (17) Nanchang. *Wu*: (18) Ningbo, (19) Shanghai, (20) Suzhou, (21) Wenzhou. *Hakka*: (22) Liancheng, (23) Meixian. *Min*: (24) Fuzhou, (25) Taiwan, (26) Xiamen, (27) Zhangping, (28) Chaozhou. *Yue*: (29) Guangzhou, (30) Yangjiang. Cheng (1991) sampled dialects (1, 2, 3, 4, 5, 6, 7, 8, 14, 17, 20, 21, 23, 24, 26, 28, 29 and 30). Wang (2004) sampled dialects (1, 5, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27 and 29).

is defined by the Xiang and Gan central areas. Mandarin itself is subdivided geographically into northern, eastern and southwestern groups. There has been a tendency among Chinese dialectologists to view these groups as more or less independent, and substantial research has been focusing on their description. However, when it comes to the question of how they are interrelated, synthetic figures are rarely given, and when they are given, they rely on formal methods constraining them to tree-likeness (Cheng 1991; Ogura 1994; Wang 1996; Wang & Wang 2004). However, diglossia, combined with massive and stratified borrowing, is expected to counter-balance the tree-like differentiation pattern by bringing genetically distant dialects into greater similarity. Given these two factors of deviation from the pure tree model, exploring the relationships between Chinese dialects in a Stammbaum framework is, *a priori*, inadequate.

In this paper, we represent Chinese dialects' relationships with a method that is unprejudiced with respect to tree-likeness, called neighbour-net (Bryant & Moulton 2004). This method is not polarized to produce a pure hierarchical structure but assesses the relative contribution of tree-likeness and network-likeness in the structure of relationships between objects, for example, languages. This method has been applied to Indo-European, and confirmed the tree-likeness of

development of this family (Bryant *et al.* 2004). Here, we apply it to Chinese dialects with the aim of retracing their trend of development.

2. MATERIALS AND METHODS

Tree reconstruction methods come with several indices that evaluate how well the data fit onto the reconstructed tree. Nevertheless, none can evaluate explicit alternatives to the pure tree model (Bryant *et al.* 2004), that is, to a connected graph with no cycles. Alternatively, the neighbour-net approach belongs to a family of methods that do not assume such an *a priori* model. It is an agglomerative procedure similar to the neighbour-joining method, but which allows the graphs to contain cycles, and thus to be locally network-like. The neighbour-net generates splits graphs from pairwise distances between the taxa (objects under study). A split is a partition of the set of taxa into two non-empty subsets. When all possible splits are computed over a set of taxa, they can either be compatible or incompatible with one another. In the first case, there is a single way of connecting the taxa, which is a perfectly tree-like branching pattern. In the latter case, there are multiple ways of connecting the taxa, resulting locally in a network. The neighbour-net summarizes the branching parts (edges) and the local

networks (boxes) in a single graphical representation. Edges correspond to single unambiguous links while boxes embody conflicting connections. Neighbour-nets were constructed using the SPLITSTREE 4B06 software (Huson & Bryant 2004).

The data supporting this analysis is lexical. The relevance of using the lexicon in phylogenetic investigations is hotly debated. Indo-Europeanists, for instance, consider that regular sound changes and innovations in inflectional morphology are better informants of linguistic descent than vocabulary, which is comparatively sensitive to borrowing (Ringe *et al.* 2002; Balter 2003). This is not the pre-eminent view in Chinese. Chinese dialectologists and austronesianists place much importance on lexical evidence (Sagart 1993; Ogura 1994; Chen 1996). Over the time span covering the formation and evolution of Chinese dialects, Ogura (1994) argues that the lexicon is more useful than other linguistic features because it can cover a time-span of several centuries/millennia, whereas vowels, consonants, tones, morphosyntax and semantics exhibit changes over smaller time periods. In this study, we used two types of lexical data. The first type of data is controlled sampling, which is assumed to be more informative about phylogeny; the second type is deemed a better representative of the whole lexicon. In each case, dialect sampling ensures that each major family is represented.

The first dataset is a character matrix corresponding to the compilation of the Swadesh 200 wordlist lexical items for 22 synchronic Chinese dialects (Wang 2004) and for Old Chinese. The Shanghai dialect (SH) is present under two different descriptions, bringing the number of taxa to 24. In what follows, we refer to Old Chinese (OC) as a dialect for the sake of simplicity, although it has a different time span to the other sampled dialects and represents the state of language spoken at about 600 BC. The 200 wordlist is a test-list that was introduced by Swadesh (1952) to support a lexically based model of language change. Such test-lists of meanings are constructed upon the assumption of their cultural universality. They are now referred to as Swadesh lists, and contain putative cultural universals such as body parts, lower numerals, topographical terms, kinship terms, personal, demonstrative and interrogative pronouns, naturally occurring phenomena and basic activities. These meanings are also assumed to belong to the basic core vocabulary, which is assumed to be more immune to change, more retentive and less subject to borrowing during language contact events than the rest of the lexicon. The 200 wordlist proposed by Swadesh is adapted to Chinese following Chen (1996). Swadesh (1955) also defined a 100 wordlist, which is a subset of the 200 wordlist. We have also considered a supposedly more conservative subset of the 200 wordlist, containing 35 items and proposed by Yakhontov (quoted in Starostin 1991 p. 59–60). It is assumed that as the list grows larger, it becomes less conservative and less immune to borrowing.

The wordlists were recoded into cognate sets. Two languages are said to be cognate for a given meaning if the forms of the lexical items in each of them can be traced back to a common root. Each meaning defines a character, and the cognate sets of a given meaning define the character states. Two distances were computed: (i) the Hamming distance, which is defined as the number of characters bearing different states between two dialects, and (ii) the lexicostatistical distance (Swadesh 1952), which is computed as minus

the logarithm of the proportion of shared cognates between two dialects.

The second dataset we have used is a distance matrix derived from similarity indices computed by Cheng (1991) on the *Hanyu Fangyan Cihui (Cihui)* lexical database. This database lists 905 lexical items in Putonghua and their variants at 18 sites of the Chinese dialectal domain (figure 1).

Cheng (1991) computed Pearson's correlation coefficients for each pair of dialects to measure their degree of closeness. Ogura (1994) defines the lexical distance based upon these measures of similarity as minus the logarithm of the correlation coefficient, in analogy to the calculus of the lexicostatistical distance. This distance is directly comparable with the lexicostatistical distance, and allows the studying of the impact of lexicon sampling on the inferred patterns of interrelation of Chinese dialects. Figure 1 shows the sites of the dialects sampled in each dataset and their linguistic affiliation according to the traditional classification of Chinese dialects by Yuan (1980).

3. RESULTS

(a) *Trees and neighbour-nets*

Preliminary tree reconstructions on the basic wordlists using distance-based (neighbour-joining; Saitou & Nei 1987) and maximum-parsimony character-based methods (PAUP 4.0b10; Swofford 1998) resulted in trees that, to some extent, agreed with the traditional classification of these dialects. Whereas the data fit onto the most parsimonious branching patterns was good (consistency index > 0.8), the bootstrap showed poor support for all clades. Bootstrap is a permutational test that assesses the level of support of the data for the individual clades on the tree. Such results suggest that the tree model is inadequate to represent the relationship between Chinese dialects. This is confirmed by our neighbour-nets, which show a pattern of relationships that strongly deviate from the pure tree, with a large amount of conflict between all dialects (figures 2–4). There is no significant difference in neighbour-net structure depending on the distance used. This was expected because these two distances are highly correlated. A feature common to all neighbour-nets, regardless of the type of lexicon and wordlist used, is their geographical tripartition into north (Mandarin and Wu), central (Xiang and Gan) and south (Hakka, Yue and Min) groups. This geographical discrimination power of our results coincides with the sharp dialect boundaries acknowledged by Chinese dialectologists, which corroborates the relevance of the figures obtained by our neighbour-net analysis.

(b) *Neighbour-nets on basic lexicon*

All three lists display distinguishable dialectal groups virtually united by a web of conflicts (figures 2 and 3). The expected Gan, Xiang, Min and Hakka are recovered for all lists. There is much conflict between and within Mandarin and Wu, which do not cluster for the 35 and 100 wordlists (figure 2). For the 35 wordlist, the Eastern Jianghuai Mandarin dialects (Yingshan, Wuhan) cluster with their geographical neighbours Xiang and Gan, but do not cluster with their putative genetic northern and southern Mandarin relatives. This group clusters closer to the other

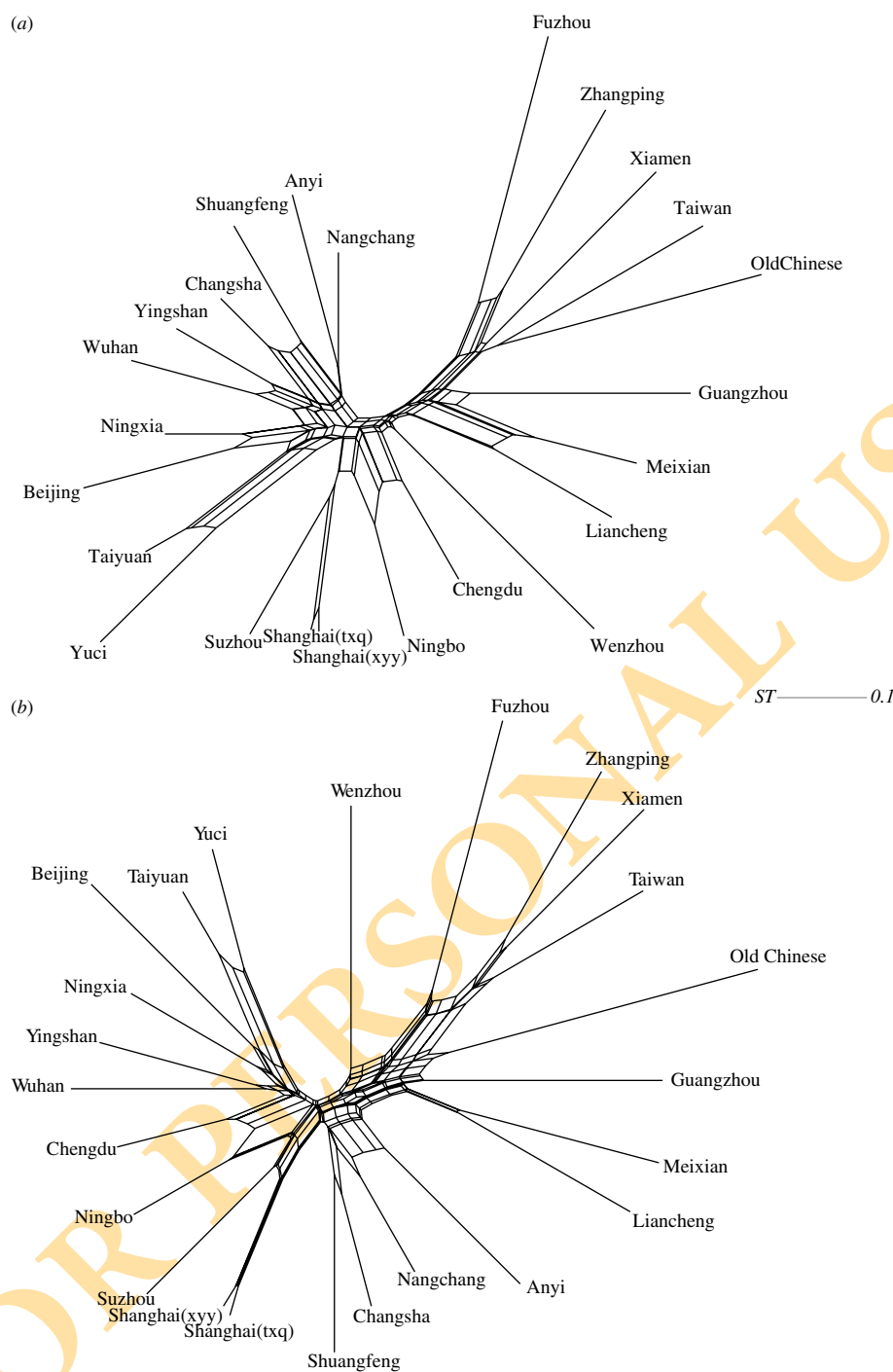


Figure 2. Neighbour-nets for the (a) 35 and (b) 100 basic wordlists (Hamming distance). Strong conflict appears at the basis of distinctive dialect groups.

Mandarin dialects when the 100 wordlist is used, but the Chengdu southern dialect then fails to define a distinct Mandarin group. The two dissident groups, Mandarin and Wu, finally cluster for the 200 wordlist, but they are still poorly supported, the edges defining them being relatively short.

(c) *Neighbour-nets on non-basic lexicon*

The neighbour-net computed on non-basic vocabulary, using the similarity indices of Cheng (1991), shows a figure similar to those obtained for basic lexicon (figure 4). It also displays a sharp split between north-central dialects (Mandarin, Xiang and Gan) and southern

dialects (Min, Hakka and Yue), with an ambiguous intermediate position of Wu dialects. In fact, the Wu group is virtually split into a north-south partition. The Suzhou dialect, representing northern Wu, is attracted towards the Mandarin group whereas the Wenzhou dialect, representing southern Wu, is attracted towards the southern Hakka, Yue and Min groups.

4. DISCUSSION

(a) *Deviation by borrowing or non-tree-like development?*

We would have expected the neighbour-nets obtained from the basic lexicon to have a different structure from

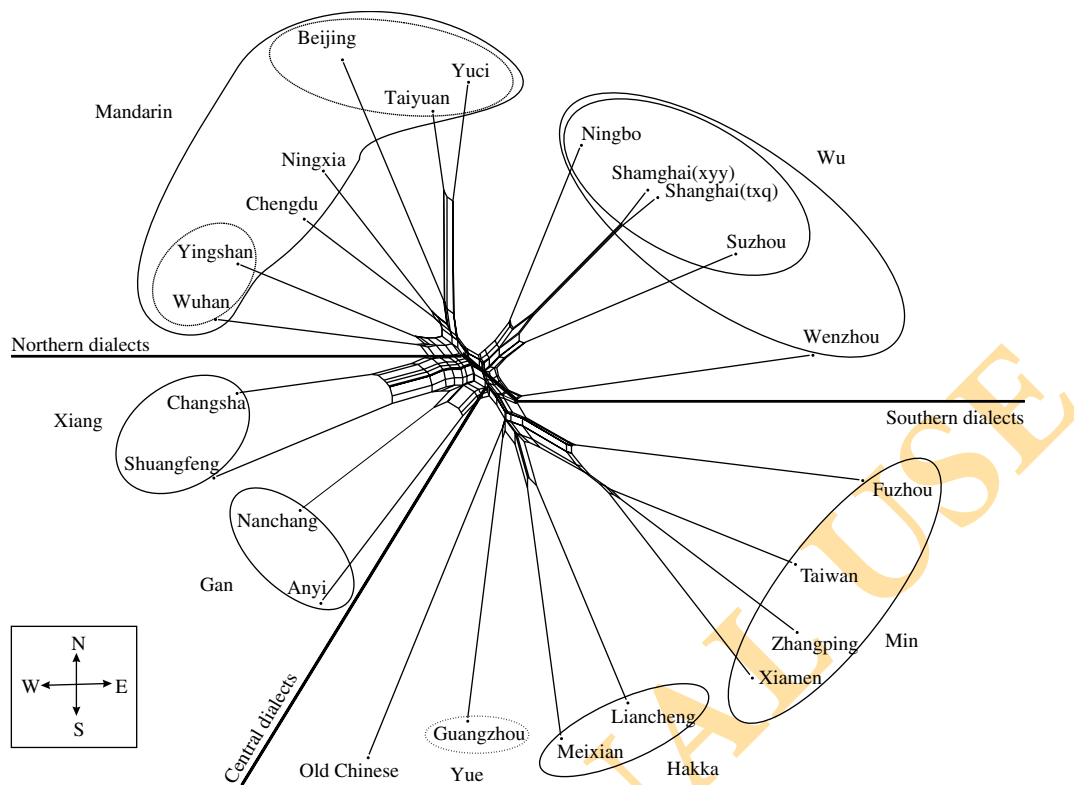


Figure 3. Neighbour-net for the 200 basic wordlist (Hamming distance). There is a strong conflict at the basis of distinctive dialect groups, where relative positioning is roughly geographical.

the one obtained from the non-basic sample. However, all neighbour-nets display the exact same pattern of differentiation, with huge conflict at the basis of distinct dialect groups, irrespective of the size and type of the sample used. A possible explanation of the absence of the expected sampling effect is that borrowing between Chinese dialects was so intense that it ‘contaminated’ even the subsets that are supposed to be more immune to it. The validity of this claim can be checked using the relative behaviour of Min with respect to OC as a reference for how a conservative data sample should behave. In fact, Min is known to have numerous archaisms, as it has conserved retentions from its ancestral form, supposedly a variety of OC (Norman 1988). The more conservative the list, the more affinity there is supposed to be between Min and OC, as Min is very retentive and OC reflects an old state of language. This is the case because OC clusters *within* the strongly supported Min group when the 35 wordlist is used, then distinctly and as a sister-group to it when the 100 wordlist is used. The distinction between OC and Min becomes even sharper with the 200 wordlist. This suggests that the 35 wordlist is indeed very conservative. Moreover, this supports the assumption that the shorter the list, the more conservative it is. Consequently, the claim that massive borrowing operating *a posteriori* of dialect divergence is the pre-eminent factor of deviation from the pure tree model should be rejected.

The alternative explanation is that Chinese dialects did not develop in a pattern of tree-like speciation. Our neighbour-nets suggest a continuum of dialects, rather than a sequence of sharp splits into distinct circumscribed entities. The differentiation of these dialects did not proceed through a tree-like speciation pattern, as speech communities were never isolated from each other due to diglossia

and intensive contact. In fact, their development was subject to two opposing forces: (i) a centrifugal force generated by the spread of people and of their languages to more distant regions, which created increasingly greater differences between the spoken varieties of an original language, (ii) in contrast, diglossia had a centripetal influence, bringing the regional varieties into greater uniformity around the standard canon. The neighbour-nets clearly portray these two opposing influences: a central conglomerate of conflict virtually unites all dialects but still sketches distinct subgroups highly in agreement with the consensual classification. They also suggest that these counterbalanced forces have acted throughout the history of China (Norman 1988), since both basic and non-basic vocabularies display the same figure.

(b) Strong dialect boundaries, genetics and geography

Little is known about the nature of dialect boundaries in China (Norman 1988). In fact, the demarcation between dialects varies depending upon the data considered—phonology, grammar or lexicon. Norman suggests that it is more appropriate to speak of strong versus weak boundaries than in terms of absolute demarcation.

The demarcation between Min and its neighbours is the strongest one displayed by our graphs, especially the one with Wu, which is in agreement with Norman’s expectations. Min varieties are likely to have also experienced a secluded development, as was the case for technology and economy, despite multiple massive influxes from the north (Zhou 1991). In fact, Fujian province, the homeland of Min speaking people, is isolated from nearby areas owing to a ragged terrain. This would have contributed to the divergence of Min from the neighbouring groups.

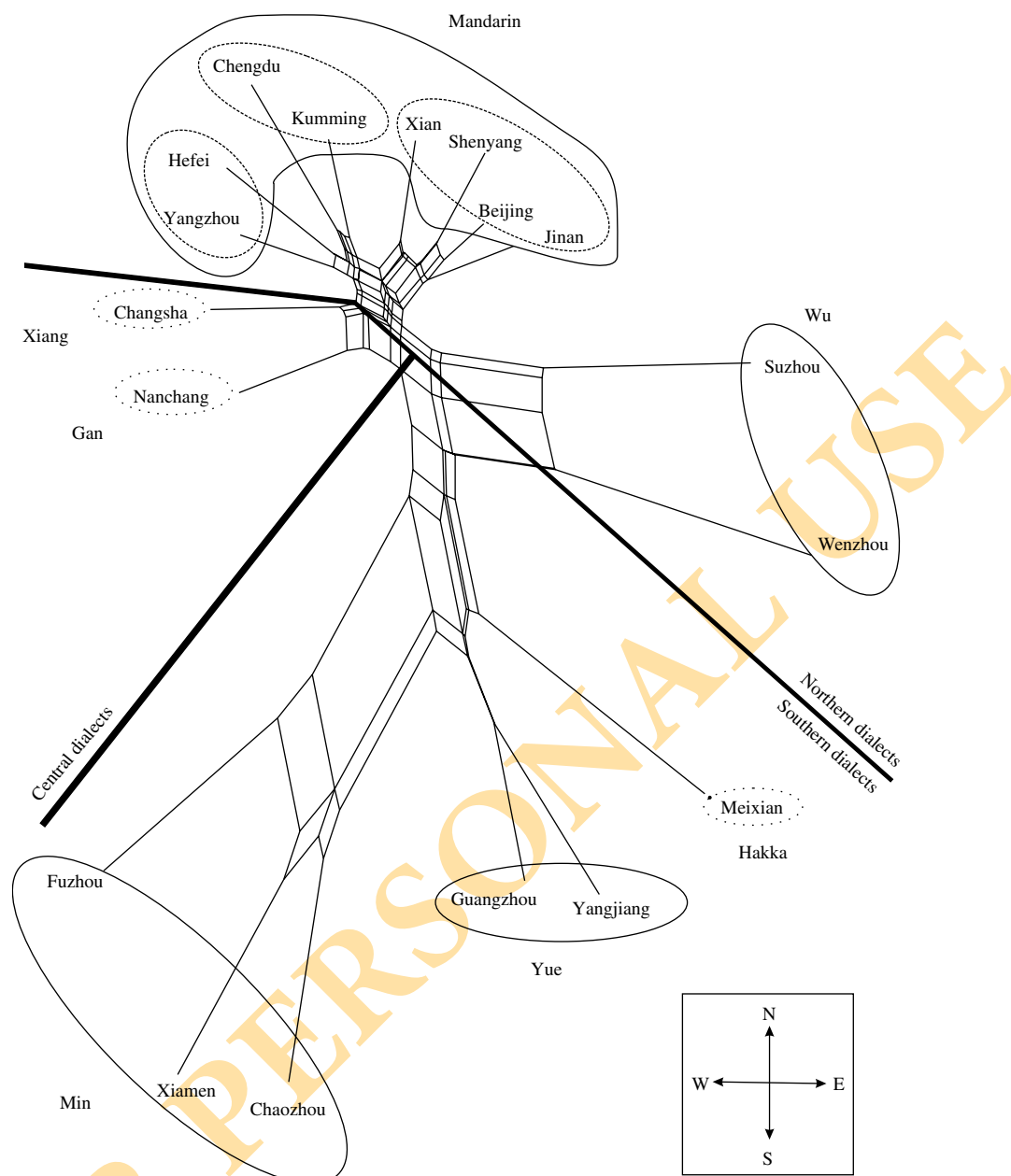


Figure 4. Neighbour-net for the distance matrix computed from Cheng (1991) similarity indices. The scales of figures 3 and 4 are the same. As for the graphs in figure 3, there is a strong conflict at the basis of distinctive dialect groups.

The initial founding effect and the later isolation, associated with a reduced population density, have probably boosted its idiosyncratic development. The phonological idiosyncrasies characterizing this group (Norman 1988) appear to be further supported by lexicon.

The second strongest boundary recovered by our neighbour-nets is located between Hakka and Gan, which supports Norman's claim that these two groups have very fundamental differences in the area of basic vocabulary. Gan has been suggested to have a privileged link with Hakka (Luo 1950) on the basis of phonological arguments. In contrast, Norman (1986, 1988) favours a closer relationship between Hakka and the other southern (Min and Yue) dialects. As suggested by Yuan (1980), Gan's lexicon is not sharply differentiated from the dialects and resembles Wu and Xiang more than it resembles Hakka. However, Sagart (2002) emphasizes that

the Hakka migrations recorded in the period between 1550 and 1850 accounts for an increased similarity in lexicon between Hakka, Yue and Min in relation to Gan. Leong (1997) shows that migrations from the Hakka heartland occurred along three axes: northeast into Fujian Min lands, south into the Yue-speaking domain, and northwest in Gan highlands. As a side effect, these migrations replaced old boundaries by new ones, blurring the pattern of genetic affiliation between the southern and central groups, especially at the level of the lexicon, which is particularly sensitive to population contact. As for Hakka, it displays tight relationships with Yue, despite a sharp demarcation. This is in agreement with the history of Hakka people, who settled during the middle of the Tang dynasty (618–907) in areas already occupied by Yue-speaking populations (Zhou 1991). The graphs also show close links between Hakka and Min, which are assumed to

have shared an early period of common development that can be traced back to the second and third centuries BC (Norman 1988).

All neighbour-nets display at least a north–south geographical partition of China. A central group is sketched out with the 100 wordlist, dividing China into northern (Mandarin and Wu), central (Gan and Xiang) and southern dialects (Yue, Min and Hakka). This pattern is maintained when larger wordlists are used, although the demarcation between northern and central zones becomes slightly blurred, which supports the claim that the central zone is transitional (Norman 1988) because it displays specificities of both northern and southern dialects. Neighbour-nets also display a neat dichotomy between northern Wu and the Wenzhou southern dialect, corresponding to the division of Wu into northern (Jiangsu) versus southern (Zhejiang) types, as suggested by Chao (1967). Wenzhou is attracted by Min whereas northern Wu dialects (Ningbo, Shanghai and Suzhou) are attracted by Mandarin, which reaffirms Yuan's (1980) claim that southern Wu is very different from northern Wu in vocabulary and that Mandarin is a strong influence on northern Wu (Chao 1967; Norman 1988). Wu is viewed as the convergence of various influences from the neighbouring Mandarin, Gan and Min (Norman 1988). This explains the improvement of Wu's clustering as the wordlist grows longer while putatively more homoplastic.

(c) *Weak dialect boundaries and the legacy of China's demic history*

Our neighbour-nets do not convey the expected strong boundary characterizing Mandarin (Norman 1988). In fact, Mandarin is virtually related to all dialects and especially to those spoken in its bordering areas. Both the central zone and Wu attract Mandarin, which tends to shorten the edges supporting Mandarin subgrouping and thus to weaken the support for this group. The figures agree with the role played by Mandarin in the linguistic and migratory history of China. In fact, not only is the standard language a Mandarin variant, but the main migrations also started in the northern Mandarin-speaking regions and headed southwards, through the central zone. A marked lexical heterogeneity is observed within Mandarin. This can be explained by the fact that the Mandarin domain is vast and was itself subject to recurring migrations. In fact, borrowing can render genetically distant dialects similar, but it can also increase the dissimilarity between genetically related ones, since the dialects display differentiated borrowing behaviours.

It is believed that Xiang was more strongly influenced by northern dialects (Norman 1988). The influence of Mandarin can be suspected in the weak demarcation between Xiang and Mandarin on the neighbour-nets. It may also explain the reduced distinction between the central group and Mandarin in terms of non-basic vocabulary. The weakness of Wu's influence on Xiang when basic wordlists are used may be related to the massive affluxes of northern settlers in the Jianxi Hakka-speaking region in the middle of the Tang dynasty (618–907). In fact, Zhou (1991) suggests that this migratory wave was associated with an economical boost, which, in turn, triggered a population boost. The language of the migrants in Jianxi would have acted like

a divider, permanently separating Wu and Xiang, and confining Min to Fujian. However, the supposedly strong influence of Wu is especially obvious when non-basic lexicon is used, and can be visualized by large boxes connecting Wu and Xiang.

5. CONCLUSION

In summary, neighbour-nets of Chinese dialects are highly consistent with geography, linguistic knowledge and China's demic history. Chinese dialects are interrelated in a figure that suggests that two opposite forces are at work: a dispersion force related to demic movements and internal change, and a homogenization force related to diglossia and heavy borrowing. Neighbour-nets allow the visualization of a model of development for these dialects that trees could not account for. They also display known reticulations between dialects, which can be reinterpreted in the wider framework of the dynamics of speech communities. Major migratory events can be retraced on the neighbour-nets, suggesting a rich linguistic legacy of China's demic history. The figures obtained from basic vocabulary are similar to those obtained from non-basic lexicon, which is supposed to be more sensitive to borrowing and representative of the whole lexicon. This stability of lexicon sampling suggests that these forces have long been at work during the development of these dialects.

The use of distance-based approaches in a phylogenetic framework is hotly debated, as a global similarity approach does not distinguish between similarities that result from common ancestry and similarities that result from horizontal transfers (such as borrowing), universal features or convergence. This is a fundamental problem when trees are generated from distances, since these conflicting evolutionary trajectories are expected to fit into a single path connecting the taxa. The neighbour-net approach avoids this limitation by not constraining the representation to an *a priori* polarized model. Therefore, the neighbour-net approach has further explanatory power. However, it is not clear how inferences made from these graphs could be validated (Bryant *et al.* 2004), where, to date, the significance of a split is obtained by external validation only.

This research was funded by the CNRS-OHLL program (France), and supported in part by grants to William S.-Y. Wang from the Research Grants Council (Hong Kong) and from Academia Sinica (Taiwan). The author thanks Wang Feng from the Language Engineering Laboratory, City University, Hong Kong, P.R.C., for providing data and advice.

ENDNOTE

¹clade: (gr. clados) putative group defined by an ancestor and all its descendants in a phylogeny.

REFERENCES

- Balter, M. 2003 Early date for the birth of Indo-European languages. *Science* **302**, 1490–1491.
 Bryant, D. & Moulton, V. 2004 Neighbor-net: an agglomerative algorithm for the construction of phylogenetic networks. *Mol. Biol. Evol.* **21**, 255–265.

- Bryant, D., Filimon, F. & Gray, R. D. 2004 Untangling our past: languages, trees, splits and networks. In *The evolution of cultural diversity: phylogenetic approaches* (ed. R. Mace, C. J. Holden & S. Shennan). UCL Press.
- Chao, Y. R. 1967 Contrastive aspects of the Wu dialects. *Language* **43**, 92–101.
- Chen, B. 1996 *Lun yuyanjiechu yu yuyanlianmeng*. Beijing: Yuwen chubanshe.
- Cheng, C.-C. 1991 Quantifying affinity among Chinese dialects. In *Journal of Chinese Linguistics Monograph Series: Languages and Dialects of China* (ed. W. S.-Y. Wang), pp. 78–112. Berkeley, CA: University of California.
- Dixon, R. M. W. 1997 *The rise and fall of languages*. Cambridge: Cambridge University Press.
- Ferguson, C. A. 1959 Diglossia. *Word J. Linguist. Circle N Y* **15**, 325–340.
- Huson, D. & Bryant, D. 2004 SplitsTree. See http://www-ab.informatik.uni-tuebingen.de/software/jsplits/welcome_en.html
- Leong, S. T. 1997 *Migration and ethnicity in Chinese history*. Stanford: Stanford University Press.
- Luo, C. 1950 *Yuyan yu wenhua*. Peking: Beijing Daxue.
- Norman, J. 1986 *What is a Kejia dialect?. Second international conference on sinology*. Taiwan: Taipei Academia Sinica.
- Norman, J. 1988 *Chinese*. Cambridge: Cambridge University Press.
- Ogura, M. 1994 Dialect formation in China: linguistics, genetic and historical perspectives. In *In honour of William S-Y. Wang: interdisciplinary studies on language and language change* (ed. M. Y. Cheng & O. J. L. Tzeng), pp. 349–372. Taipei, Taiwan: Pyramid Press.
- Ringe, D., Taylor, A. & Warnow, T. 2002 IndoEuropean and computational cladistics. *Trans. Philol. Soc.* **100**, 59–129.
- Sagart, L. 1993 Chinese and Austronesian: evidence for a genetic relationship. *J. Chinese Linguist.* **21**, 1–63.
- Sagart, L. 2002 *Dialect variations in Chinese. Papers from the third international conference on sinology*. pp. 129–153. Taipei: Institute of History and Philology, Academia Sinica.
- Saitou, N. & Nei, M. 1987 Neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**, 406–425.
- Schleicher, A. 1853 *Die ersten Spaltungen des indogermanischen Urvolkes. Allgemeine zeitung fuer wissenschaft und literatur*.
- Schmidt, J. 1872 *Die Verwandtschaftsverhältnisse der Indogermanischen Sprachen*. Weimar: H. Böhlau.
- Starostin, S. 1991 *Altajskaja Problema i Proisxoždenie Japonskogo Jazyka* [The Altaic problem and the origin of the Japanese language]. Nanka, Moscow.
- Swadesh, M. 1952 Lexico-statistic dating of prehistoric ethnic contacts. *Proc. Am. Philos. Soc.* **96**, 453–463.
- Swadesh, M. 1955 Towards greater accuracy in lexicostatistic dating. *Int. J. Am. Linguist.* **21**, 121–137.
- Swofford, D. L. 1998 *PAUP: Phylogenetic analysis using parsimony (and other methods)*. Sunderland, MA: Sinauer Associates.
- Wang, W. S.-Y. 1996 Linguistic diversity and language relationships. In *New horizons in Chinese linguistics* (ed. C.-T.J. Huang & Y.-H.A. Li), pp. 235–268. Dordrecht: Kluwer Academic Publishers.
- Wang, F. 2004 BCD: basic-words of Chinese dialects <http://chinese.pku.edu.cn/wangf/wangf.htm>
- Wang, F. & Wang, W. S.-Y. 2004 Basic words and language evolution. *Lang. Linguist.* **5**, 3.
- Yuan, J. 1980 *Hanyu fanyan gaiyao*, 2nd edn. Beijing: Wenzhi Gaige Chubanshe.
- Zhou, Z. 1991 Migrations in Chinese history and their legacy on Chinese dialects. In *Journal of Chinese Linguistics Monograph Series: Languages and Dialects of China* (ed. W. S.-Y. Wang), pp. 29–51. Berkeley, CA, University of California.

As this paper exceeds the maximum length normally permitted, the authors have agreed to contribute to production costs.