PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B

Phil. Trans. R. Soc. B (2010) **365**, 3829–3843
doi:10.1098/rstb.2010.0099

Splits or waves? Trees or webs? How divergence measures and network analysis can unravel language histories

Paul Heggarty[1,*], Warren Maguire[2] and April McMahon[2]

[1]*Linguistics, Max Planck Institute for Evolutionary Anthropology, Deutscher Platz 6, 04103 Leipzig, Germany*
[2]*Linguistics and English Language, University of Edinburgh, Dugald Stewart Building, 3 Charles Street, Edinburgh EH8 9AD, UK*

Linguists have traditionally represented patterns of divergence within a language family in terms of either a 'splits' model, corresponding to a branching family tree structure, or the wave model, resulting in a (dialect) continuum. Recent phylogenetic analyses, however, have tended to assume the former as a viable idealization also for the latter. But the contrast matters, for it typically reflects different processes in the real world: speaker populations either separated by migrations, or expanding over continuous territory. Since history often leaves a complex of both patterns within the same language family, ideally we need a single model to capture both, and tease apart the respective contributions of each. The 'network' type of phylogenetic method offers this, so we review recent applications to language data. Most have used lexical data, encoded as binary or multi-state characters. We look instead at continuous distance measures of divergence in phonetics. Our output networks combine branch- and continuum-like signals in ways that correspond well to known histories (illustrated for Germanic, and particularly English). We thus challenge the traditional insistence on shared innovations, setting out a new, principled explanation for why complex language histories can emerge correctly from distance measures, despite shared retentions and parallel innovations.

Keywords: tree; network; phylogeny; historical linguistics; language history; language divergence

1. SAME LANGUAGES, DIFFERENT VISIONS

There is no clearer way to illustrate the topic of this paper than by contrasting three different representations of the sub-grouping relationships within the same language family: Indo-European. Figure 1 reproduces one of Ringe *et al.*'s (2002, p. 90) 'best trees' from their search for a 'perfect phylogeny' for the family. Figure 2 reproduces, also for Indo-European, the 'consensus tree' produced by Gray & Atkinson (2003, p. 437) out of a sample of 1000 possible configurations. Figure 3, meanwhile, is a NeighborNet analysis of Indo-European based on a distance matrix derived from the binary values inherent in the 'isogloss map' of the family by Anttila (1989, p. 305), also reproduced here as figure 4. For more on how this NeighborNet was produced, including Anttila's specification of his data characters, see the electronic supplementary material, www.languagesandpeoples.com/Eng/SupplInfo/AnttilaNeighborNet.htm.

The contrast could hardly be clearer: 'trees' in figures 1 and 2; a 'web' or network in figure 3. The first two are structured entirely by binary splits; in the third there is almost no such branching, and the relationships between the subgroups take the form of a network of *cross-cutting* relationships instead.

Most importantly, the difference is by no means merely one of representation, but has implications for our understanding of the relationships between the early Indo-European languages and the real-world context in which they arose. For what moulded the particular constellation of relationships between the dialects and languages of any family was none other than the unfolding relations between the populations who spoke them, as they themselves diverged through (pre-)history. The opposing linguistic patterns in figures 1–3 therefore also imply contrasting visions of what must have happened in the real world, during the early divergence history of Indo-European, to account for how those patterns came about.

Yet there was only one real-world population history, of course. So how is it that for this same language family, different types of phylogenetic analysis can come to such radically different outputs? Is one right and the others wrong? And how might the different types of phylogenetic method—tree-only or network approaches—help us uncover what actually happened in linguistic prehistory?

2. LANGUAGE DIVERGENCE AND THE REAL WORLD: TWO MODELS, ONE REALITY

Linguists have traditionally represented patterns of divergence within a language family in terms of either of two discrete models by which they are assumed to arise:

- the splits model, corresponding to a branching family tree structure;

* Author for correspondence (paul.heggarty@gmail.com).

One contribution of 14 to a Theme Issue 'Cultural and linguistic diversity: evolutionary approaches'.

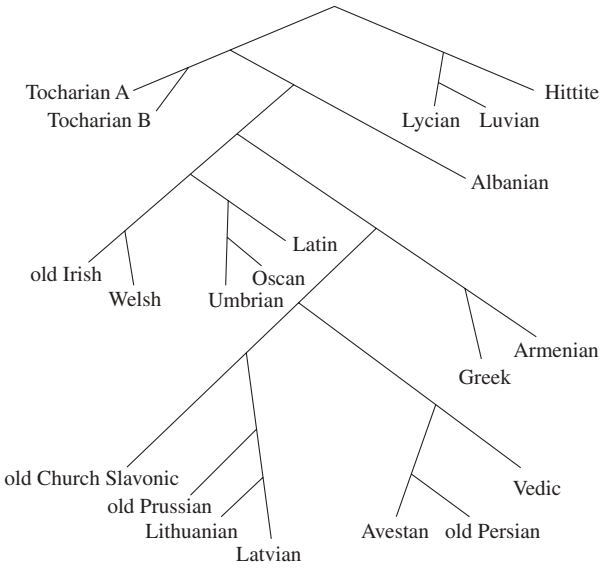

Figure 1. 'One of the best trees with Germanic omitted'. (Reproduced with permission from Ringe *et al.* 2002, p. 90.)

- the 'wave' model, typically yielding a (dialect) continuum.

There has been much debate in theory as to the respective merits and demerits of these models, and how well suited each might be for representing the types of relationship that can obtain between language varieties within a family. What has been all too briefly considered, however, is how and why *in practice* these two types of relationship come to arise in the first place. How and why should it be in the nature of languages to develop into relationships of these particular and starkly contrasting types?

In truth the contrast in models only exists at all because it reflects two very different processes in the real world. Broadly speaking, these two mechanisms are as follows.

- A speaker population divides, prototypically by long-distance migration(s), into two (or more) groups, henceforth physically separated from one another. This leads to the classic language split pattern.
- A speaker population expands (whether suddenly or more progressively) over a continuous territory, across which a degree of contact is maintained—at least at the immediate local level, and all the more strongly the shorter the distance between any two points. This leads to language divergence in a pattern of overlapping, cross-cutting waves. Cases would include, but are certainly not limited to, expansions by the 'demic diffusion' model (Renfrew 1989, pp. 126–131).

The nature of any given language family as either more split-like or more wave-like can thus be interpreted as effectively a linguistic record of the past, pointing to one or other of these two different mechanisms as the probable history of its speaker populations, the real-world scenario that moulded that family's particular pattern of divergence.

Let it be clear from the outset, then, what the nature of the relationship is between the divergence pattern of a language family and the real-world contexts in which its speakers lived. For this relationship is unambiguously one of cause-and-effect; and in a direction that is equally ineluctable. Whether, how, and which particular languages diverge is not just some natural law of 'what languages do' or 'how languages evolve'. Forces in the real-world context—demographic, socio-political, cultural, and so on—are the cause; they alone determine entirely the linguistic *divergence* effects.

One must hasten to clarify that what those forces do *not* determine is the form and nature of whatever particular language changes arise (other than in cases of contact). Changes can be highly idiosyncratic, and generally are either random, or in line with other changes in the language system as a whole (Heggarty 2006, p. 188). Changes *arise* by natural linguistic processes, then; what external forces determine is only whether those changes (whatever linguistic form they take) either develop independently and differently, or come to be shared, from one region to the next. That is, real-world forces dictate not which particular *changes* occur, but the patterns of language *divergence* that they ultimately give rise to. It is in real-world forces that lies all the difference between on the one hand, the 'singletons' within Indo-European, such as Albanian and Armenian; and on the other hand, the great families like Romance, Germanic or Slavic, born out of vast expansions propelled by the might of Rome and the turmoil that attended its fall.

What many readers may feel is missing from the above discussion of the two divergence mechanisms is the role of borrowing or contact. Certainly, when faced with data on language relationships that cross-cut in ways incompatible with branching trees, as a stock explanation to rescue a tree-only analysis its advocates typically roll out a 'splits-then-borrowing' model, seen as a more tree-friendly alternative to the dialect continuum. It is with very good reason, however, that we leave borrowing out of this section. For there must be no blurring of the distinction between splits-then-borrowing and 'dialect continuum' scenarios: the two represent very different visions of real-world language histories. For a start, borrowing and contact are modes of language *con*vergence (of languages either unrelated, or related but already diverged from each other). As such, neither has any real place here, in this discussion of the two basic modes of language *di*vergence out of a common ancestor.

The contrasts in figures 1–3 pose a conundrum for the prehistory of Indo-European in particular, then. How could those three studies come to such radically different results? Which of the two basic mechanisms of language divergence comes closest to what actually happened to the peoples who spoke the earliest Indo-European languages? Answering those questions calls for a full exploration of the difference between a splits-then-borrowing and a dialect continuum scenario, and how significant it is for how we understand and model language prehistory. These issues have to be left for a more wide-ranging survey than there is space for here, however, in further papers by Heggarty (in preparation *a*,*b*).

The treatment here will be limited to only one of the main issues in the trees versus webs debate, and seek instead to establish a more general point of methodology. We start out from a first way to defuse the potential stand-off between the tree and wave models, which is to progress beyond a simplistic vision that sees the two as mutually exclusive either/or alternatives for any one language family. For in reality, neither a branching tree nor a continuum model alone is sufficient to account for the complex relationships observed across many a language family. Nor, indeed, have we any reason to expect either to be. The vagaries of history typically ensure that in the real world, both types of process may act upon the populations speaking any given language family, combining in any manner of ways across time and space. The real-world history of any language family need by no means be a story of uniquely one or the other mechanism, but is very often a complex composite of the two.

This complexity has implications for the tools and data we might look to in order to model and represent relationships between the languages within a family. In principle, to do justice to real language divergence histories, we need a model able to capture both split and wave mechanisms within a single analysis and representation. Indeed ideally we would wish for a model that allows us to tease apart the respective contributions of each to the overall story.

3. NETWORK METHODS: BEST OF BOTH WORLDS?

The ability to do just this is precisely the claim made for one particular type of phylogenetic analysis: those of the network type, in contrast to others of the tree-only type. Though initially developed for applications in the biological sciences, particularly genetics, two network-type methods in particular have also been widely applied to language divergence data: Network and NeighborNet, both of which we shall survey here.

There is in fact another well-known network-type analysis method, namely Split Decomposition by Bandelt & Dress (1992), now integrated into the SplitsTree 4 package (Huson & Bryant 2006). It is not considered in detail here, however, firstly because it has been rather less used in linguistic studies of late. A second and more critical objection is that for precisely the task in hand here, that of teasing apart the respective strengths of tree-like and web-like signals, Split Decomposition has an inherent bias—towards the former. As McMahon & McMahon (2005, p. 158) put it, with Split Decomposition, 'graphs based on bigger and more complex datasets tend to become more tree-like by default'.

It is not the task of this short article to set out in detail the workings of these methods. Useful general sources include the valuable and readable survey of and introduction to network methods, as applied specifically to language studies, in Bryant *et al.* (2005), and discussion and illustrations for more language families in McMahon & McMahon (2005). Here, we just briefly overview how the two network-type methods we cover here have been received in historical linguistics, and focus on how they relate to the issue of particular interest in this paper.

(a) *Network*

The Network algorithm (Bandelt *et al.* 1995, 1999) was developed by a group led by the mathematician Hans-Jürgen Bandelt and geneticist Peter Forster. The programme can be downloaded from www.fluxus-engineering.com, and for a brief explanation of how the algorithm produces its network outputs, see Forster *et al.* (1998, pp. 182–184). For the purposes of this paper, one of the key defining characteristics of Network, in contrast to NeighborNet, is that it takes as its input format individual state data, not overall measures of distance or similarity between languages (see §3*c*).

Forster in particular has applied Network to language data, in papers with various colleagues. Forster & Toth's (2003) study of Celtic languages, however, was rather taken to task by historical linguists. Criticisms of their approach to the language data and certain assumptions in the dating methodology employed are widely felt to invalidate the paper's conclusions. Rather less problematic are Forster *et al.* (1998) on Alpine Romance varieties, and Forster *et al.* (2006) on Germanic. These too, though, are based on rather limited datasets. The authors start out from the Swadesh list of 100 basic word-meanings, but for many of these the data are invariant across all language varieties in the study, or missing, or cause 'chaotic reticulation'. The authors remove all of these data, which in the case of Germanic leaves an effective dataset of just 28 data-points. Moreover, each of these is open to the established criticisms of traditional lexicostatistics as to the bluntness of its 'all-or-nothing' binary approach to how related languages overlap in their lexical semantics. Questions also remain as to the authors' inferences for what their outputs may mean for the history of the Germanic-speaking populations (Forster *et al.* 2006, pp. 135–136).

Still, however one might question the handling of the data and ancillary assumptions in any one case, such objections are besides the main methodological point for the purposes of this paper. For criticism on these scores does not impugn the algorithm *per se* as one that in principle does harbour considerable potential as a means of representing language divergence relationships. One does not have to agree with all aspects of Forster and his colleagues' approach in order to grant that their Network algorithm does indeed offer one means of differentiating, 'weighting' and combining in a single representation both the tree-like and web-like components of the overall complex of relationships within a language family. As an illustration, figure 5 shows its output for the Germanic languages, reproduced here from Forster *et al.* (2006, p. 134). Albeit on an imperfect and limited dataset, the pattern does duly reflect that cross-cutting relationships exist within Germanic—whether imputable to shared or parallel innovation, to cross-cutting waves across a dialect continuum, or to contacts

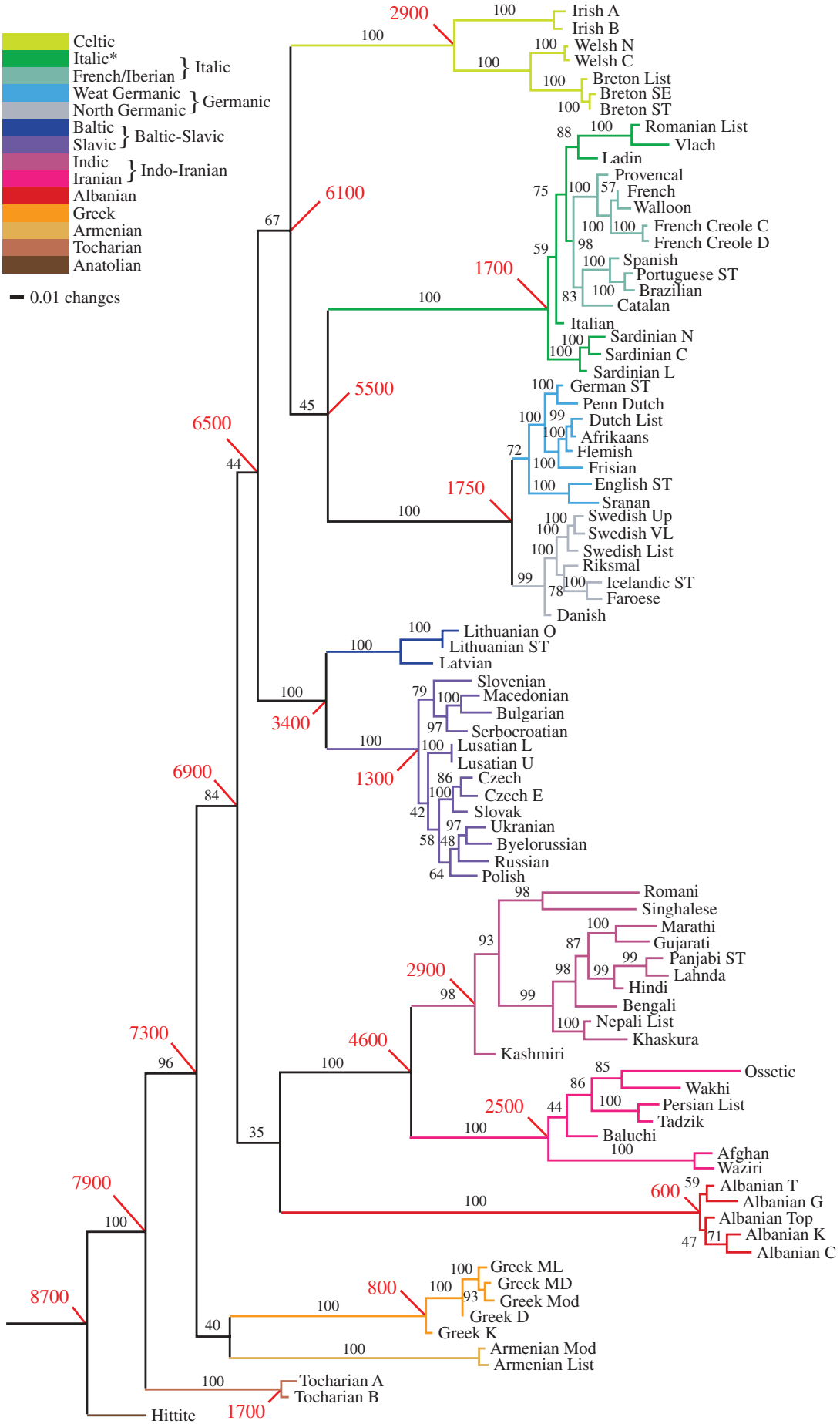

Figure 2. Consensus tree for Indo-European (Reproduced with permission from Gray & Atkinson 2003, p. 437). The numbers up to 100 along each branch (in small font) are posterior probability values—see text. The numbers up to 8700 at each node (in large font) are time-depth estimates expressed in years BP.

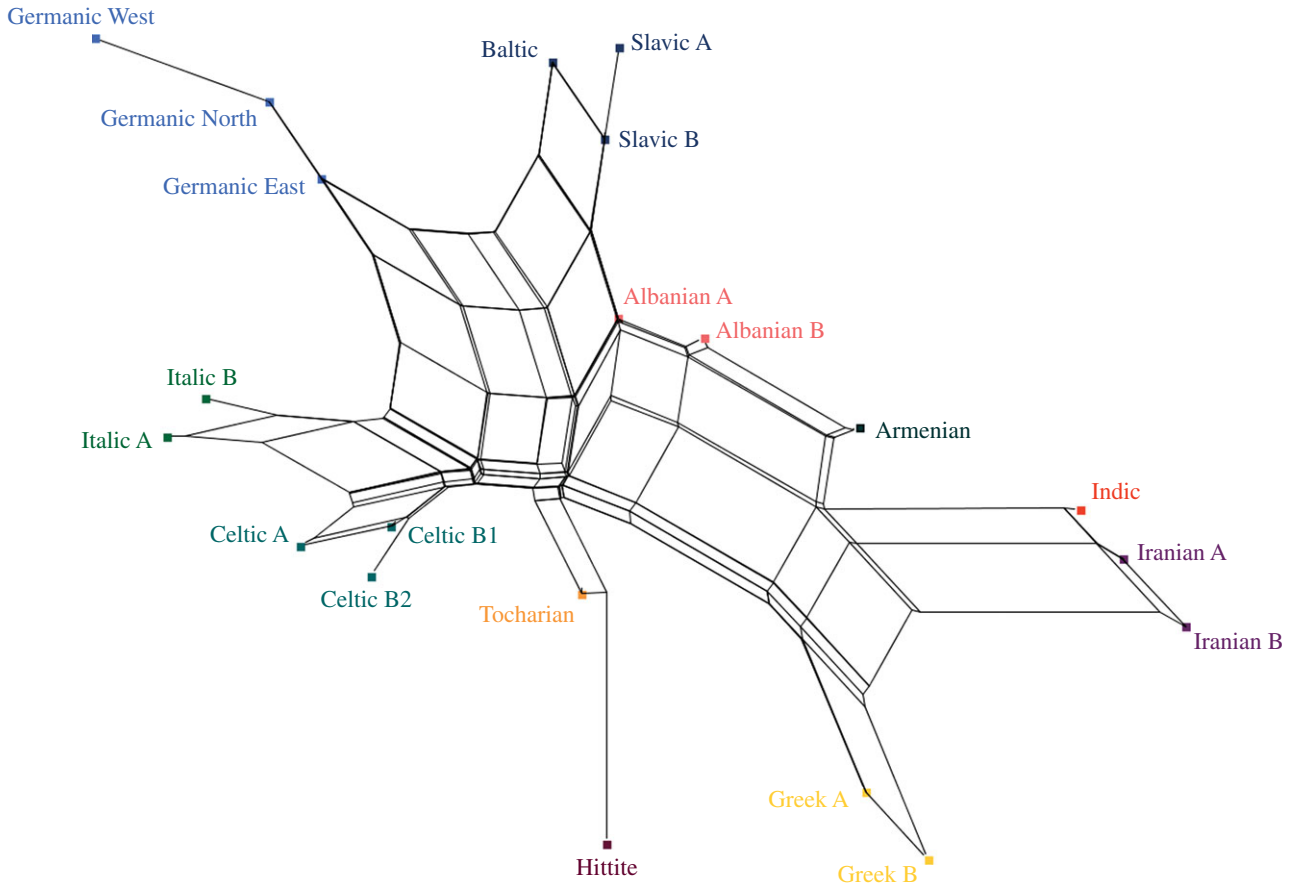

Figure 3. NeighborNet of a distance matrix from Anttila's (1989, p. 305) Indo-European isogloss map (figure 4).

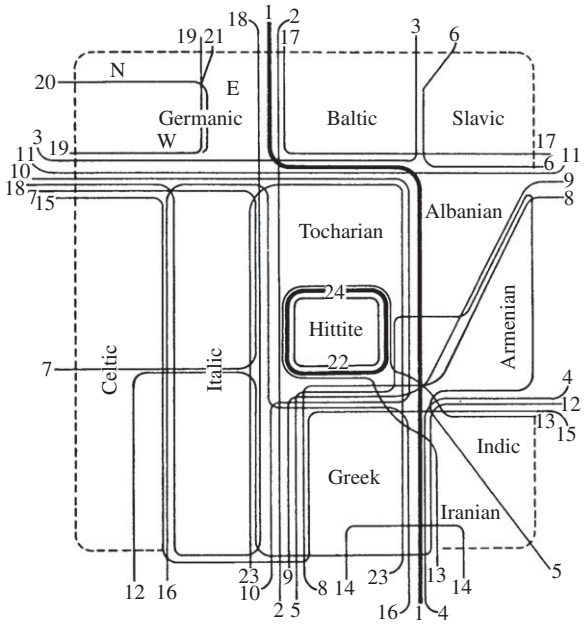

Figure 4. A dialect map of the Indo-European languages. (Reproduced with permission from Anttila 1989, p. 305.)

between speakers after an earlier split. Such an output *format* is arguably more realistic and balanced than a tree-only representation. It also has the attraction (and advantage over NeighborNet) of identifying all individual changes in the network diagram itself, i.e. in figure 5 the word-meanings in which a cognate change is 'reconstructed'.

For other authors' applications of Network to language data, see McMahon & McMahon (2005, pp. 140–154).

(b) *NeighborNet*

NeighborNet, of which figures 3, 6 and 7 are illustrative outputs, was developed by Bryant & Moulton (2004), and is now integrated into the SplitsTree 4 package (Huson & Bryant 2006). It too was first intended particularly for applications in the biological sciences, but has been enthusiastically advocated for applications to language data by a number of researchers, including April McMahon and Russell Gray, each together with various colleagues. Unlike Network, NeighborNet takes as its input format not state data, but overall measures of distance between languages (see §3*c*).

Bryant *et al.* (2005) provide an instructive talk-through of the process by which the method goes about turning its input data into its output representation, as applied to illustrative language data. This includes an application to Indo-European (not, of course, the *tree* reproduced here in figure 2), while Holden & Gray (2006) apply it to Bantu. In both studies, the authors use traditional lexicostatistical data.

Other researchers have also used NeighborNet on traditional lexicostatistical data, applied to Australian and Indo-European languages in McMahon & McMahon (2005, pp. 164–165), for example. We have also applied it to other, finer-grained measures of language divergence. Heggarty's study of the

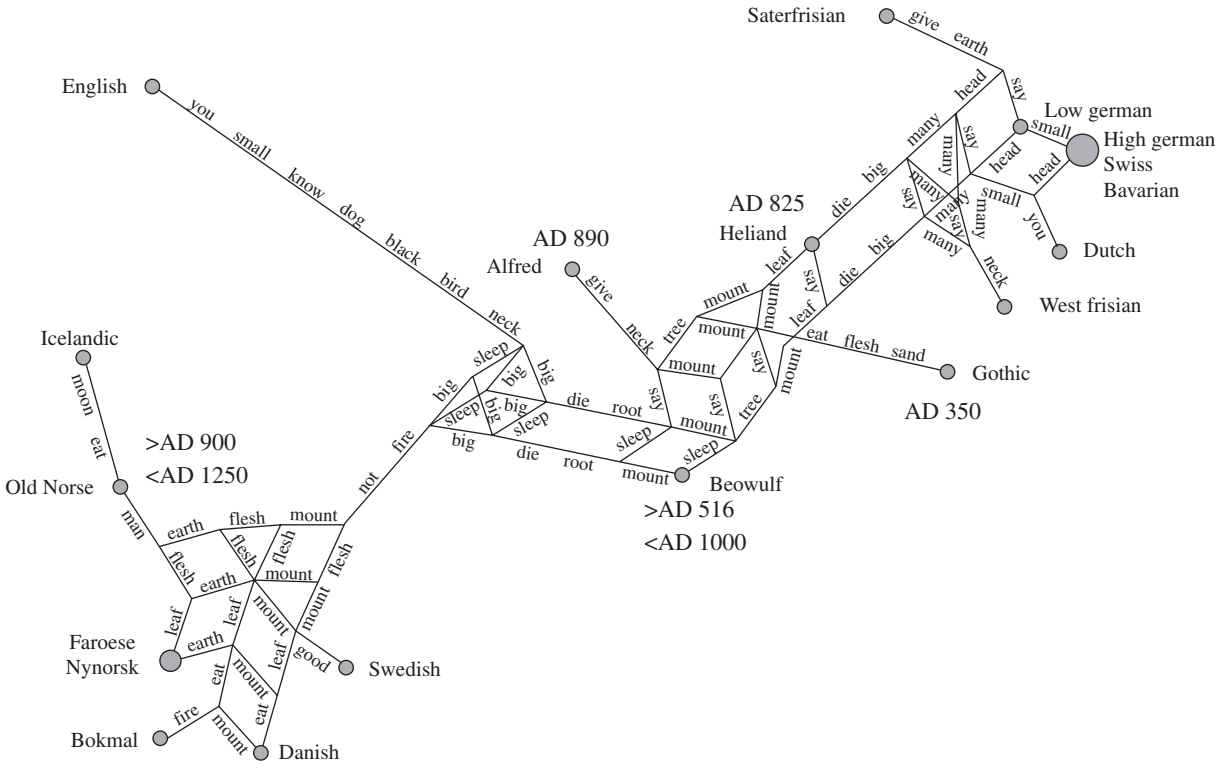

Figure 5. Unrooted network of 19 Germanic language samples. (Reproduced with permission from Forster *et al.* 2006, p. 134.)

Andean language families Quechua and Aymara is based on distance data in lexical semantics, as calculated by a new quantification method to work to a more refined level than traditional lexicostatistics, set out in Heggarty (2005), and more briefly in McMahon *et al.* (2005) and McMahon & McMahon (2005, pp. 166–173). We have explored NeighborNet also with divergence ratings in *phonetics*, for the Romance languages (Heggarty *et al.* 2005), and for accents and dialects of English and other Germanic languages (see figures 6 and 7 later; McMahon *et al.* 2007; Maguire *et al.* 2010).

(c) *Measuring network versus tree signals*

Both of these network-type methods, then, are able to visualize together the relative strengths of the tree-like and web-like signals within a single dataset. This is indeed how Forster *et al.* (1998, p. 174) explicitly present their Network method, while Bryant *et al.* (2005, p. 67) make much the same claim of NeighborNet. As applied to language data, both can effectively represent the split and/or wave elements within the overall divergence pattern of a family. For the purposes of uncovering prehistory, the value of this balance is as a record of the particular mix of the two main real-world processes that underlie the tree-like and web-like sectors of that overall pattern: physical separations, including migrations (yielding branching patterns); and/or expansions over continuous territory (leaving wave patterns).

Moreover, as well as providing a way of *representing* this balance graphically, network-type methods can also be seen as a means of effectively weighing up, indeed *measuring* the overall 'tree-ness' or 'net-ness' of a particular dataset. Together the various network-type methods provide a range of approaches to assessing this numerically, such as the 'splittable percentage' ratings produced by the Split Decomposition network method (Bandelt & Dress 1992). For NeighborNet, meanwhile, see in particular the 'delta scores' discussed in Gray *et al.* (2010).

In fact, even methods whose graphical outputs are only in tree format can nonetheless produce similar quantifications of tree-ness, many of them based on samples of large numbers of possible trees. As examples, in their study on Indo-European Gray & Atkinson (2003, p. 436) report that 'a preliminary parsimony analysis produced a consistency index of 0.48 and a retention index of 0.76'—effectively, measures of how well the data fit on the tree. (There are, however, known problems with consistency indices particularly, with actual scores overly dependent on sample size.) One can also focus on individual branches within the tree. In figure 2, Gray & Atkinson's consensus tree includes a 'posterior probability' value specified above each branch, which can stand as an indication of how strongly supported it is across the set of possible trees—themselves valuable data on Indo-European prehistory (Heggarty in preparation *a*, §§4.1, 4.2.3, 6).

Other recent work takes existing tree-only approaches as a basis upon which to build what are effectively new forms of network analysis. Dealing with large sets of possible trees allows such collections to be synthesized into a consensus network, for which an algorithm has been devised by Holland & Moulton (2003). Atkinson & Gray (2006, p. 97) illustrate this for Indo-European, though they see it in this case as 'just [a] useful pictorial summary of the [...] fundamental output', namely the *distribution* of possible trees that is their core interest for their proposed dating methodology. Nakhleh *et al.* (2005, pp. 399–400),

Table 1. Comparative lexical data, by cognate set, for four basic meanings in five Romance languages.

	EAT	SLEEP	GO	HOUSE
Portuguese	*comer*^{A}	*dormir*^{A}	*ir*^{A}	*casa*^{A}
Spanish	*comer*^{A}	*dormir*^{A}	*ir*^{A}	*casa*^{A}
French	*manger*^{B}	*dormir*^{A}	*aller*^{B}	*maison*^{B}
Italian	*mangiare*^{B}	*dormire*^{A}	*andare*^{B}	*casa*^{A}
Romanian	*mânca*^{B}	*dormi*^{A}	*merge*^{C}	*casă*^{A}
Cognate Sets *derived from different Latin variants*				
A	*comedere*	*dormīre*	*īre*	*casam*
original sense	eat up	sleep	go	hut, cabin
B	*mandūcāre*	—	*ambulāre*	*mansiōnem*
original sense	chew		walk about	place to stay
C	—	—	*mergere*	—
original sense			immerse, go under	

Table 2. A table of multi-state data, derived from the data in table 1.

	multi-state datum 1 (EAT)	multi-state datum 2 (SLEEP)	multi-state datum 3 (GO)	multi-state datum 4 (HOUSE)
Portuguese	A	A	A	A
Spanish	A	A	A	A
French	B	A	B	B
Italian	B	A	B	A
Romanian	B	A	C	A

meanwhile, have developed an extension of their original perfect phylogeny approach that yielded figure 1, seeking to produce now perfect phylogenetic networks, which they explore for Indo-European—though they still present their networks essentially as trees, albeit with contact edges.

In practice, network-type methods have often been used just as initial diagnostic tests, to judge whether a particular dataset yields a pattern that is tree-like 'enough' in order to justify proceeding to tree-only analyses useful for particular research ends. It was on the strength of this sort of test, for instance, that Gray & Atkinson concluded that the Dyen *et al.* (1992) lexicostatistical dataset for Indo-European yields a fairly strongly tree-like signal, and thus felt confident in going on to use a tree-only analysis for their approach to dating the family.

And yet, as the contrasts between figures 1–3 here suggest, different datasets and different methodological approaches can nonetheless lead to very different assessments of how tree-like or web-like was the divergence of the same family. For Ringe *et al.* and Gray *et al.*, their data and analyses give a pattern that is basically tree-like for Indo-European, while the NeighborNet from Anttila's data suggests quite the contrary. Much of the explanation for this apparent contradiction lies in whether one considers the entire family, throughout its history, or focuses on just its *earliest* divergence stages. Many later developments in Indo-European were obviously 'tree-like': all modern Romance languages clearly derive from a single Proto-Romance ancestor, all Germanic languages from Proto-Germanic, and so on. The impact is to make the overall signal for the family more tree-like, by 'diluting' what appears to be the more web-like structure of its earliest divergence—the same characteristic that has for so long frustrated attempts to establish an agreed higher-order branching for the family. This issue is taken up more fully in Heggarty (in preparation *a*, §4.1).

(d) *Data formats: multi-state, binary or distance measures?*

We turn now to the key difference between Network and NeighborNet for our purposes in this paper, namely in the format of the input data each takes. Network takes multi-state data; NeighborNet takes distance measures.

To illustrate how these represent two very different approaches to the same language data, tables 1, 2 and 3 provide an example for four word-meanings (EAT, SLEEP, GO and HOUSE) in five Romance languages (Portuguese, Spanish, French, Italian and Romanian). The 'raw' language data can be set out as in table 1, giving each language's principal lexeme in that meaning, and identifying which of the 'cognate sets' for that meaning across Romance that lexeme falls into, symbolized here by the letters A, B, C, and so on.

By a cognate set is meant a collection of words (technically lexemes), one from each of several different languages within a family, which all derive directly (i.e. without borrowing) from the same original lexeme in that family's common ancestor language—even if sound changes since may have left them rather different in precise phonetic form. For the EAT meaning, for example, word-forms in the Romance languages fall into two main cognate sets:

- the set of word-forms derived from Latin *comedere* (originally *eat up*), including Spanish *comer* and Portuguese *comer*, cognate with each other (and spelt identically, though somewhat different in pronunciation);
- the different set derived instead from Latin *mandūcāre* (originally *chew*), including French *manger*, Catalan *menjar*, Italian *mangiare* and Romanian *mânca*, also all cognate with each other.

Table 3. A triangular matrix of pairwise distances between languages, derived from the data in table 1.

Portuguese	Spanish	French	Italian	Romanian	
—	$^{0}/_{4}$	$^{3}/_{4}$	$^{2}/_{4}$	$^{2}/_{4}$	Portuguese
	—	$^{3}/_{4}$	$^{2}/_{4}$	$^{2}/_{4}$	Spanish
		—	$^{1}/_{4}$	$^{2}/_{4}$	French
			—	$^{1}/_{4}$	Italian
				—	Romanian

Table 4. A table of binary data, derived from the data in table 1.

	binary datum 1 = EAT^{A} *comedere*	binary datum 2 = EAT^{B} *mandūcāre*	binary datum 3 = SLEEP^{A} *dormīre*	binary datum 4 = GO^{A} *īre*	binary datum 5 = GO^{B} *ambulāre*	binary datum 6 = GO^{C} *mergere*	binary datum 7 = HOUSE^{A} *casam*	binary datum 8 = HOUSE^{B} *mansiōnem*
Portuguese	1	0	1	1	0	0	1	0
Spanish	1	0	1	1	0	0	1	0
French	0	1	1	0	1	0	0	1
Italian	0	1	1	0	1	0	1	0
Romanian	0	1	1	0	0	1	1	0

For the SLEEP meaning, meanwhile, the major Romance languages offer just one major cognate set, since all use cognates derived from Latin *dormīre*. For the meaning GO there are three sets: cognates of Latin *īre*, *ambulāre* and *mergere*; e.g. Spanish *ir*, French *aller* and Romanian *merge*, respectively. Table 1 illustrates such patterns of shared cognates for our four sample word-meanings in five sample Romance languages.

The patterns in table 1 can be converted into a table of multi-state data as in table 2. This is the input format used by Network, in which the particular cognate state for each meaning in each language remains distinguished. Any single change in cognate state in any language can be reconstructed along all edges of the network, as per the meanings attached to them in figure 5.

An alternative approach to the same raw language data is to convert them instead into a triangular grid or matrix of 'pairwise' distances. With lexicostatistical data such as these, distance measures are typically calculated by simply counting the proportion of the (here, four) meanings in the list for which a given pair of languages use lexemes that are *not* cognate (e.g. Spanish *comer* but Italian *mangiare*), relative to the meanings for which their lexemes *are* cognate (e.g. Spanish *casa* and Italian *casa*). These ratings could equally well be expressed as 'similarity' rather than distance ratings, by simply subtracting each of the results shown from 1.

Once converted in this way, it is no longer possible to recover from these conflated overall distance measures the individual data that underlie them, i.e. to tell apart in which particular meanings any two given languages do or do not share cognates. So unlike Network, or even isogloss maps as in figure 4, there is no way for NeighborNet itself to identify which particular aspects of its output diagrams correspond to which particular linguistic differences. (Although in practice this can often be inferred by other means: see for instance the discussion of the rhotic versus non-rhotic division among varieties of English in McMahon *et al.* 2007, pp. 136–137.)

This has often been raised as an objection to the utility of distance measures in linguistics, and of methods such as NeighborNet that rely on them—the key issue we shall consider in the remainder of this paper. Certainly, the difference between state and distance data is a crucial one that also distinguishes our three Indo-European representations here: figures 1 and 2 are both based on state data; figure 3 on distance data.

Finally, it is important to note that even methods which use state rather than distance data fall into two quite different sub-types, since state data may be either

- multi-state: i.e. more than two states for a given datum, such as A, B and C for multi-state datum 3 in table 2 (the meaning GO); or
- binary: i.e. only two states per datum, presence versus absence, or 1 versus 0.

(The Network package, in fact, initially had only an algorithm for binary data, i.e. Bandelt *et al.* (1995). Another was later included which takes multi-state data, i.e. Bandelt *et al.* (1999)).

Again, the same raw language data can be analysed in either of these two ways. With traditional lexicostatistical data as in table 1, the two possible approaches differ in terms of the question(s) asked of the word-form registered in each language, for every meaning in the Swadesh lists.

- A single information question: into *which* cognate set—A, B, C, D, etc.—does this language's word-form fall? This yields one *multi-state* datum per meaning in the list, with all values taken as equally different from one another.
- A series of yes/no questions, one for each cognate set: *is* cognate A present or absent in this language? This yields one *binary* datum (yes or no) for each of

the various cognate sets found for this meaning across all languages.

Table 3 has already illustrated multi-state analysis by word-meanings in the list; table 4 now shows the alternative binary analysis by cognate sets. In this latter case, each of the various meanings in the list converts into a range of cognate sets, more for some meanings and fewer for others: in our Romance sample, just one cognate set for the meaning SLEEP, two for EAT, three for GO and two for HOUSE. The result across all four of these meanings combined is a total of eight cognate sets (i.e. $1 + 2 + 3 + 2$).

For the full Swadesh 200-meaning list for the 87 Indo-European languages in the Dyen *et al.* (1992) dataset, the first approach gives 200 multi-state data, one for each meaning in the list. The second approach, however, as used by Gray & Atkinson (2003) for the tree-only phylogenetic method that they prefer for their proposed dating technique, yields a total of 2449 binary data points, since across all 200 meanings combined there are a total of 2449 cognate sets. That is, on average there are over 12 cognate sets per meaning, though this masks significant differences between meanings with very few cognate sets (e.g. the numerals), and others with relatively high numbers of cognate sets. Objections have been raised to this approach and its implications: for how independent some of the cognate set data-points are from others; and for how some meanings in the Swadesh list are effectively weighted more or less than others. The authors maintain, however—in Holden & Gray (2006, p. 25), for example—that these make 'very little difference to the results', a finding whose logic Pagel & Meade (2006, appendix A) investigate and concur with.

4. MATCHING DIVERGENCE PATTERNS WITH THE REAL-WORLD CONTEXT

(a) *State versus distance data: theory and practice*

Returning to the contrast between state data and distance data, before one ventures into the relative unknown of early Indo-European divergence, we would do well to see first how distance data and NeighborNet perform in a case where the external history *is* largely known. Does NeighborNet live up in practice to the claim that within the overall divergence pattern of a single language family, it can tease apart tree-like and web-like components—and thus, according to the logic set out in §2 above, point separately to split-like 'migrations' and/or to wave-like dialect continua arising out of more continuous expansions?

We illustrate this with figures 6 and 7, our own NeighborNet outputs from a triangular distance matrix of measures of divergence in phonetics between language varieties within the Germanic family. Figure 6 covers 52 traditional regional languages and dialects within Germanic. Figure 7, meanwhile, compares a collection of 11 historical varieties of English against present-day 'traditional' local speech in 29 different regions. The recording data were collected by Paul Heggarty (for Germanic) and Warren Maguire (for English), and transcribed by Maguire (or in some cases, an expert on that particular variety).

The transcriptions underlying figure 6 can be accessed, and the modern recording data heard, at www.languagesandpeoples.com/Germanic. Likewise, the database of English varieties in figure 7 can be viewed and heard at www.soundcomparisons.com (websites by Heggarty). The method by which the divergence ratings in phonetics were calculated from these transcriptions was devised and programmed by Heggarty—see www.languagesandpeoples.com/Methods.htm and Heggarty *et al.* (2005).

Now of course, whether overall *distance* measures (particularly in phonetics) can stand as a valid or useful indicator of language *history* is in principle very much a moot point, which we come to shortly. Nor do geographical proximity or separation necessarily match with language descent. And certainly, a first impression of the NeighborNet in figure 6 is that it does not closely resemble the family tree structure by which the Germanic languages are traditionally classified. Aside from the now extinct 'Eastern' branch (Gothic) not included in figure 6, that traditional tree sees a primary North versus West split, with English classified within West Germanic, indeed inside a further 'Anglo-Frisian' sub-branch. On the other hand, the rather narrow and exclusive criteria applied in order to come to a binary branching tree are not the only way in which one can conceive of the real-world history of a language family, and do not necessarily make for an answer to precisely the question we are asking here. For our somewhat different perspective of language *speakers'* history, set out in §2, a look at our Germanic NeighborNet reveals that the split-like versus web-like patterns within it do in fact reflect real-world contexts and histories of speaker populations rather well. Varieties which since an early stage were geographically isolated from each other into essentially separate speaker communities duly emerge at opposite ends of the few clear splits in the overall pattern: all English (and Scots) varieties in the British Isles stand as a group off to the left, separated from all other ('continental') Germanic varieties by the one very sharp split running through the overall picture. To the right, meanwhile, is a second relatively clear split, between the Scandinavian varieties at the top, and those of continental West Germanic.

Within these major groups, however—that is, across the unitary territories settled by largely contemporaneous or progressive expansions—the pattern is very different. The English and Scots varieties, even very marked dialectal variants such as Holy Island (Northumbria), Buckie (a 'Doric' form of North-Eastern Scots) and Ulster Scots, all relate to each other in essentially web-like patterns with no particularly sharp splits. Across *continental* West Germanic, where traditional dialects still survive rather more strongly, the web-like picture is an even more striking reflection of a progressive dialect continuum across the entire region, incrementally proceeding in fairly close step with geography, from Flanders to the Alps. Not that the match with geography is perfect, of course; among the various reasons is one that goes back to certain limitations inherent in NeighborNet for

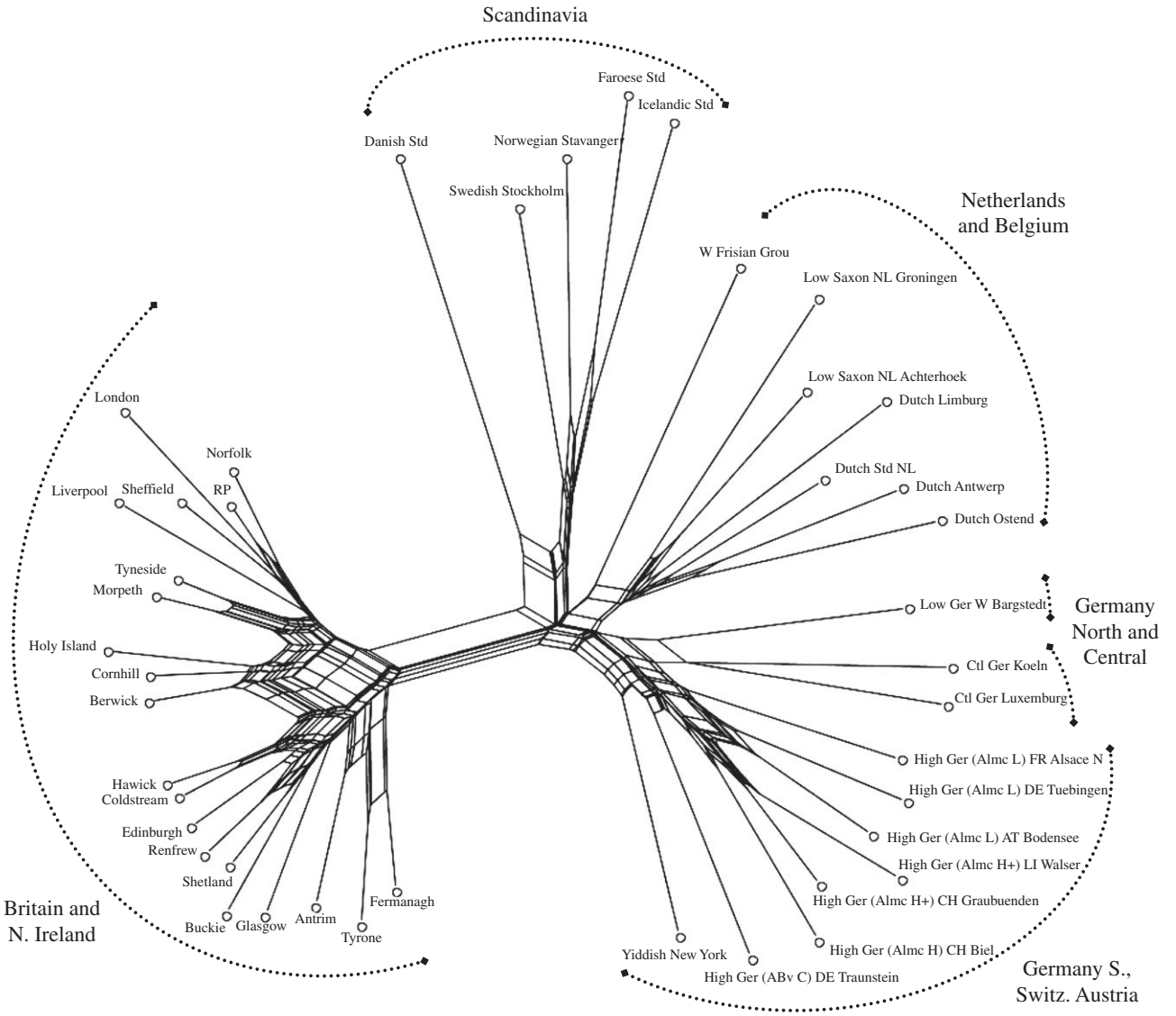

Figure 6. NeighborNet from a matrix of distances in phonetics between traditional dialects in Germanic.

representing language and particularly dialect relationships, to which we return in §4*b*. (Within Scandinavia, and over the dialect zone between southern Denmark and North Germany that links the two 'blocs', coverage in our dataset is as yet rather too sparse to draw any sound judgements. Further recordings, transcriptions and calculations for this region are underway to fill in these gaps.)

The distance-based approach used here also makes it possible to include comparisons against reconstructed historical forms, albeit with all the necessary provisos as to quite how certain and precise we can be in the phonetic transcriptions assumed for them. In this respect too, distance measures can in practice yield outputs eminently coherent with known history, as demonstrated in figure 7 which compares varieties of English from a number of different historical stages. Closest to the (nonetheless distant) Proto-Germanic, and first to 'branch off' from it, are the two Old English forms; then the four Middle English; then Early and Middle Scots. (Note that these two divisions of Scots refer to periods that actually correspond most closely to the Middle English and Early Modern English periods, respectively.)

By this stage in the history of English, clear patterns are beginning to emerge in geography too, in that while the historical Scots varieties do remain closest to the broadly contemporary English ones, vis-à-vis present-day varieties they side clearly with those of Scotland. Similarly, the positions of the Early Modern and Late Modern English of England, with respect to modern regional varieties both within Britain and overseas, reflect the respective time-depths at which English began expanding and diversifying into world-wide patterns too: first to the New World and, more recently, to the Southern Hemisphere.

These aspects of the NeighborNet can but underline a further principle critical to the correct interpretation of any distance measurements: that degrees of divergence between language varieties are a function not just of separation *time* but also of the degree of cohesiveness of a speaker community, for which geographical *space* is often a fairly close proxy, especially within a dialect continuum. This principle, and the serious consequences it entails for attempts to date language divergence from distance measurements, are explored more fully in Heggarty (in preparation *b*, §6, §7).

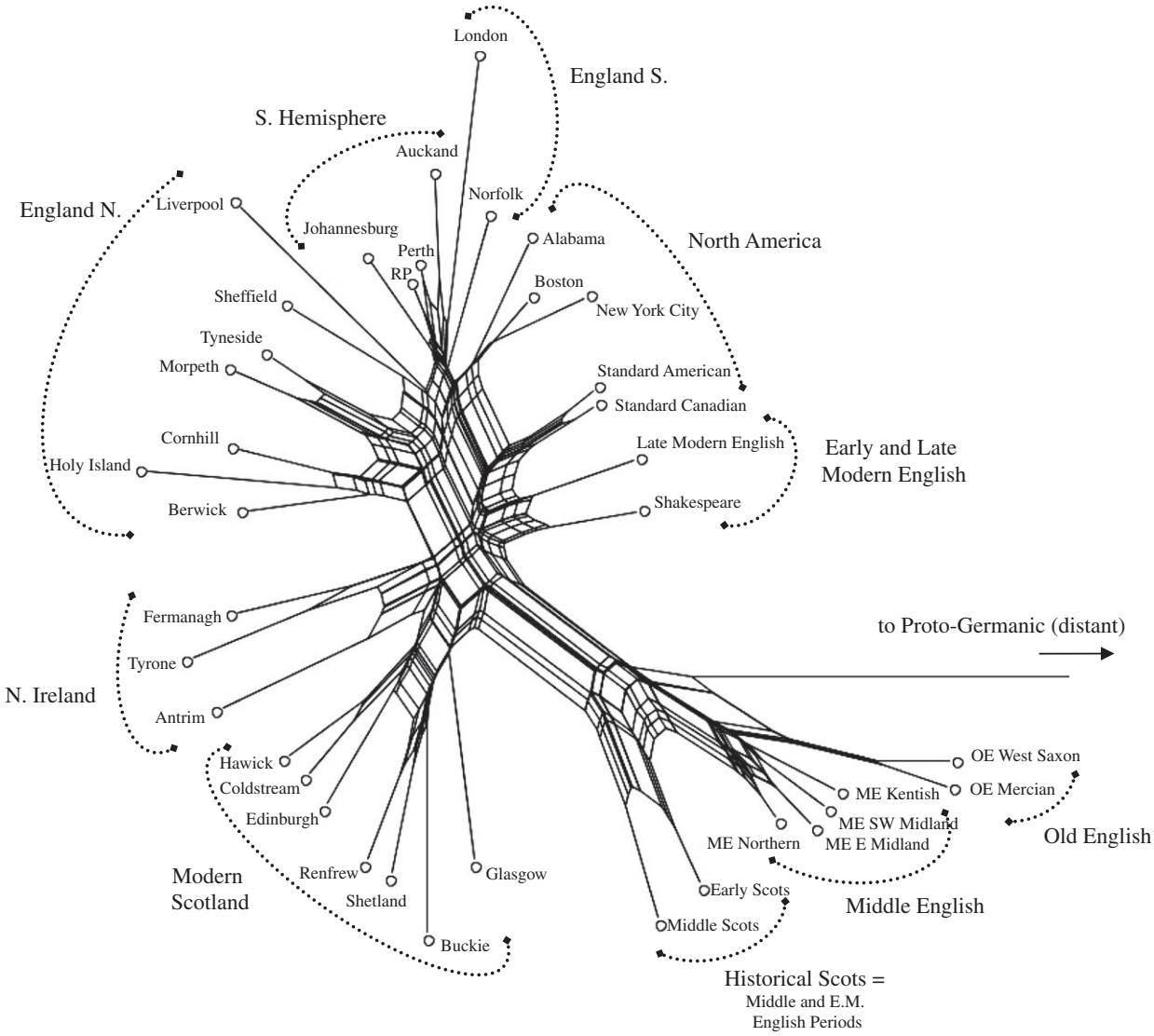

Figure 7. NeighborNet from a matrix of measures of divergence in phonetics between traditional dialects and reconstructed historical varieties of English.

(b) *Limitations of NeighborNet*

The two figures (6 and 7) also give an inkling of one potentially quite serious limitation that NeighborNet faces as a means of representing language divergence patterns, especially in dialect continua. This is that a language taxon cannot appear in the *midst* of a network of reticulations, but only around the perimeter. For in dialect continua, of course, it is not only possible but positively expected that one dialect may be intermediate between many others all around it. NeighborNet cannot place a language taxon in such an intermediate position graphically. As our coverage of English and other Germanic language varieties has expanded, it has become clear that certain intermediate language varieties can be quite unstable in where they appear in the NeighborNet outputs, depending on how one filters the wider set of language taxa. Examples include certain 'variably rhotic' varieties of English, and Frisian.

The latter is of course much debated in Germanic linguistics in any case, given how the traditional branching tree ranks it as the one variety within Germanic that is closest to English; but in the face of objections that it is certainly today closer to neighbouring varieties normally considered dialects of Dutch or of Low Saxon. The latter position is the one that emerges more strongly from these distance measures, although Frisian does also appear as something of an outlier relative to continental West Germanic, in the direction of both English (especially its rhotic varieties) and Scandinavian. In a NeighborNet which includes all of these, Frisian's position necessarily emerges as a compromise of all of these relationships (though not necessarily a happy one). If one wishes to focus only on the relationships within traditional West Germanic, one can filter out the Scandinavian varieties and this duly isolates an even stronger signal of Frisian as intermediate between continental West Germanic and English (though still closer to the former).

Furthermore, from the perspective of continental West Germanic, in a NeighborNet like figure 6 that includes 'external' varieties like English and Scandinavian, these take up the left-hand side of the network,

leaving the continental West Germanic varieties all in a group to the right. The result is that any relationships within this group are effectively limited to being expressed by the unilinear sequence they branch off in down this right-hand edge of the NeighborNet: any one variety can only be set between two others. In practice this cannot faithfully reproduce even two-dimensional geographical space on a map. Frisian, for example, would rank at a far extreme of the continuum in the north, but so too would a west Belgium variety such as Ostend. These two also need to be quite distant from one another, with other Dutch varieties in between. But those Dutch varieties are also intermediate between both of these and the Low and Central German varieties. Such triangular relationships cannot be captured in what is effectively just a unilinear sequence around the perimeter of the network, which necessarily cannot do full justice to the multi-dimensional relationships between varieties in a dialect continuum. Again, some improvement can be gained by focusing on those relationships only: if one filters out known 'external' varieties, the relationships across West Germanic then duly form a two-dimensional space within which the Frisian, Belgian and other Dutch varieties stand in rather more natural positions relative to each other. A final factor influencing outputs and stability to filtering is how many taxa are included, and how smoothly they are sampled across geographical space. Adding multiple very close varieties effectively gives greater 'weight' to their region and forces NeighborNet to treat them more stably, while varieties from less heavily sampled regions become relatively less stable to filtering.

These can be quite serious difficulties when applying NeighborNet to language data. In ongoing research we are therefore exploring alternative representation and analysis methods that do not suffer from this limitation, including multi-dimensional scaling and alternatives to traditional isogloss maps. These have their own limitations, however, and provided one is well aware of the inherent limitations and potential instabilities of NeighborNet, and careful in one's interpretations and in using filtering to focus on specific questions, it remains a powerful (and very fast) algorithm. Indeed, its limitations notwithstanding, figures 6 and 7 confirm that NeighborNet analysis can indeed distinguish split-like and wave-like patterns in the divergence of languages and dialects, in ways that can offer an impressively close match with the known external history of their speakers. And of course it achieves this from simple *distance* data (in this case in phonetics), i.e. *without* a specification of ancestral states.

(c) *'Diagnostic power', ancestral states and parallel change*

At first sight this can seem unexpected, since such data do not obey the common insistence from historical linguists that only those correspondences known to be shared innovations can inform us of language histories. Indeed, from the traditional viewpoint, an expected response to figure 3 here is that, especially once converted to distances, Anttila's data may be insufficiently reliable to be at all probative of history. The strongest view holds that correspondences between languages that reflect *shared retentions*, survivals since a common ancestor, tell us nothing, and that for any given comparative datum, if we cannot tell which of the states is ancestral, then the datum cannot be useful. Similarly, independent *parallel innovations* tell us nothing about common histories; so we should also remove from our dataset all correspondences that seem to be instances of them. The only probative characters assumed to be diagnostic of (tree) history are therefore taken to be known *shared innovations*. On these grounds, this school of thought might simply rule out some of Anttila's data.

Ringe *et al.* (2002) do not appear to sign up fully to this strongest view, though they do generally take a rather dim view of the value of distance data. Presumably not all of Anttila's data would have passed their own stringent data screening criteria; though as argued in Heggarty (in preparation *a*, §4.3.2), their screening itself is hardly without impact on their own database.

Certainly, distance data call for caution in interpreting what they can tell us of language relationships and especially histories. Nonetheless, views have been clouded by certain misconceptions as to exactly how the various phylogenetic analysis methods put state and/or distance data to use in order to generate their outputs (see Heggarty 2005, pp. 37–38, and of course the explanations of how each method actually works). More generally, other misconceptions inherent in the traditional 'strongest view', as set out above, certainly seem to overstate the limitations of distance data as informative on language history, if we are to judge from figures 6 and 7.

(d) *Why distance data work in practice: a new proposal*

For if the strong principles are so sacrosanct, how is it that our Germanic database (including the subset of varieties of English), which neither specifies ancestral states nor filters out known cases of parallel innovations, nonetheless produces NeighborNets that do indeed reflect the decisive historical factors that shaped the divergence patterns across the family? Did our study just 'get lucky' with the particular dataset in hand? Or is there in fact a stronger, principled way to account for why distance measures might reflect language history after all? I (Heggarty) would propose precisely that.

The explanation starts back at some first principles of how changes pattern across languages. At any given stage in any language's history, a vast number of changes are possible, indeed more or less 'natural'. But other than in cases of direct contact, we usually cannot explain why in any one language a given change occurred precisely when it did, nor why *that* change did occur while other possible ones did not. Or, to see it from the alternative perspective of all language lineages, the subset of them in which a given change occurs is effectively random. Consider the sound change of the vocalization of post-vocalic [l], by which (usually first via dark [ɫ]) it becomes [w], [u] or even [o]. European language varieties that

exhibit this change make for an eclectic, random collection ranging from modern Glaswegian and Estuary English (*milk* as [mɪwk]) to the Portuguese of [bɾaziw], Old French (*castle* versus *chateau* [ʃɑto]), Serbo-Croat (*Belgrade* versus *Beograd*), Polish, some accents of Bulgarian, and so on. That any individual sound change like this occurs in two or more language varieties is clearly no necessary indicator whatsoever that it happened during a period of common history, for the innovation so often occurs independently anyway, as all these known cases show. And just as the languages partaking in this parallel innovation form a subset that is essentially random, so too is its converse: the remaining subset of those languages that have *not* undergone this change, i.e. those characterized by the 'shared retention' of post-vocalic consonantal [l].

So it is both expected in principle, and the case in practice, that our dataset on phonetic distances for Germanic in part reflects changes that are not shared innovations. Examples include the parallel innovations of the vocalization/loss of post-vocalic /r/ in many English varieties and in Standard German, e.g. *here* as [hiːə] and *hier* as [hiːɐ], respectively. Likewise, they reflect the *shared retention* of ancestral dental fricatives [θ] and [ð] in Icelandic and (most varieties of) English, as in *þing*, *thing*. Like any other phonetic correspondence or difference, these register significantly in our distance data. So wherever these particular sounds occur in our database, in *this* characteristic the varieties concerned are indeed duly measured as less distant to each other than to all other varieties. How is it, then, that these parallel innovations and shared retentions do not seem to disturb the overall pattern?

The explanation is potentially twofold. Firstly, one might surmise that these parallel innovations and shared retentions must simply be far outweighed in practice by those changes that are indeed shared innovations. That may well be part of the explanation, but even to see things in terms of 'outweighing' is to fail to follow through on the first principle of language change discussed above: the effective randomness in the patterning of which language changes occur when, in which language lineages. For the very same characteristic that makes a particular language change susceptible to occurring independently in parallel also largely guarantees that it is effectively 'neutral' as to which other language varieties happen to develop it too. Indeed, the more susceptible a given change is to develop independently, the more random its distribution across language lineages should be.

So for any one independent change, the patterns of languages in which it happens to occur in parallel, and the remainder in which it does not occur, will be random. This randomness entails that taken together, all parallel changes and all shared retentions over all varieties will, in net effect, largely balance each other out. This will leave a background level of random (i.e. star-like, not net-like) divergence between language varieties, in both parallel changes and shared retentions. On top of this indistinct background, clear non-random patterns do emerge in how varieties differ from each other, of course; so what principally determines them must be the only changes that remain, namely the shared innovations. These are the changes that come to be shared in different regions not by multiple independent chance occurrences, but by being propagated, from just a single occurrence, across a wider speech community. They come to be shared, that is, not by randomness, but by the forces in the real-world context that determine the extent and nature of those communities, their degrees of coherence, and also their external boundaries that the propagation wave may not cross. In the strongest case, almost all changes will be either adopted or rejected right across a speech community, maintaining it as a single, coherent language albeit changing through time. In the weaker but arguably historically more common case, many changes will spread widely, but in waves overlapping in different patterns, leading to a dialect continuum.

It is these factors that account, in a perfectly principled way, for why—despite the contribution of parallel innovations and shared retentions to our overall distance measures—our Germanic phonetics database can nonetheless reproduce such a close match with known historical population splits, and real-world geographical patternings across the dialect continua zones. More generally, I put them forward here as a principled argument that undermines the objection that it is only identifiable shared innovations, and thus only state and not distance data, that are useful for determining ancestry by means of phylogenetic and network analysis.

On the strength of this, it transpires that distance data can in fact be perfectly relevant—provided that certain criteria are met in how they are calculated. For in order for the above principles to apply, the dataset must be a balanced, global and representative sample of the level of language in question: lexical semantics, phonetics, etc. (and ideally, of all levels together). Furthermore, the various types of linguistic difference within any one level need to be taken together and weighted for their relative significance with respect to each other, in some principled, balanced way (see Heggarty *et al.* 2005). These provisos stress once more the critical importance that attempts to put numbers on language data must attach to meaningful weightings (see also Heggarty 2006, p. 185). But if a quantification method can achieve these key requirements, its distance measures can indeed present a picture of the degrees of difference between languages which in practice can be highly informative of language history too—even if not entirely consistent with traditional branching tree representations.

Indeed, the insistence on a suitably weighted 'holistic', unfiltered database can be considered a particularly healthy aspect of distance-based approaches. For they not only allow, but ideally call for, datasets that are more complete and thus more balanced; more so, certainly, than any approach that hand-picks or applies heavy *a priori* screening or filtering to the real-language dataset. We return here to the observation in §4*a* above that figure 6 is certainly no perfect match with the traditional family tree for Germanic, for which the primary split is between North versus West, with English classified within the latter 'branch'. And yet figure 6 does show coherence with

key aspects of the real-world contexts in the population history of the speakers of Germanic languages. Those realities, as set out in §2, dictate that tree-only representations do not, indeed cannot, necessarily tell us everything significant about language histories. By prioritizing certain 'diagnostic' changes and overriding others, the tree-only approach risks misrepresenting the overall story.

If our real goal is to uncover the histories of the *populations* that spoke given languages, rather than abstract schemas intellectually satisfying for their binary purity, then it is served by using language data to arrive at a picture of the nature and degree of cohesion (or otherwise) of speech communities within a language family, through the story of its divergence. To this end, we must represent the historical and linguistic truth that English ultimately underwent a longer and more total isolation than did most continental varieties from each other. The approach that best uncovers this signal is to measure and represent the full impact of *all* those far-reaching changes that came to mark it out as so distinct from all other Germanic varieties in so many ways—rather than to limit our representation of that history simply to the earliest few isoglosses assumed to identify a initial, radical and 'exclusive' North versus West (versus East) split. It is only right that the impact of both of these different determiners of Germanic language history should show up clearly in any representation that seeks to reflect that history. Overall distance measures and network analyses are by definition better placed to achieve this balance than tree-only analyses based only on selected 'screened' data.

Furthermore, while English is widely assumed to have derived from a *mixture* of Germanic dialects—eminently logical also in terms of population history—this too cannot be represented in a tree model. That structure is inherently forced to oversimplify the most plausible history. Nor could it capture the clear possibility that of the dialectal mix that went to make up English, a greater part was of more western than northern Germanic character and provenience, but not *exclusively* so, especially at a time when the difference between those two groups may well have been very much a continuum still. It is such matters of degree, rather than mutually exclusive binary alternatives, that speak in favour of distance and network analyses for language history, and against tree-only ones. (And in the case of Germanic there are, besides these methodological objections, other grounds to criticise the traditional tree representation of its history: see Robinson's (1992, pp. 247–263) discussion of how even the same data are often interpreted very differently by different authorities.)

(e) *Lessons for Indo-European, and for historical linguistic methodology?*

To conclude, let us recall the due balance that is needed when assessing the respective utility, for uncovering language history, of distance versus state data, and of tree-only versus network-type phylogenetic methods. Certainly, our intent in contrasting figures 1, 2 and 3 here is by no means to underestimate the value of Gray & Atkinson's or Ringe *et al.*'s pioneering studies. For many purposes, the tree idealization undoubtedly has immense practical utility. To echo Atkinson *et al.*'s (2005, p. 209) point, when it comes to devising models, known lies are indeed permissible, if they are the sort that can help lead us to the truth. Among these valuable idealizations is at least the possibility of a dating mechanism as put forward by Gray & Atkinson. To be sure, questions remain as to their method, but it is precisely one of its attractions that the authors are at pains to limit the impact of the tree idealization, by 'dating' not from a single tree but from a *distribution* of many 'most plausible' trees, and their various respective time-depths (as their method calculates them).

The point here is hardly to claim, then, that distance data and network analyses are uniquely valuable, and state data and branching-tree analyses necessarily less so. But it most definitely is to redress the balance. We should check the traditional historical linguist's instinct that all data that cannot be confirmed as shared innovations are to be discarded as valueless. For an equally principled case can be made, as here, for why the supposed limitations of such data turn out to be far less serious than has generally been assumed. In our search to uncover language prehistory, we are only the poorer if we overlook the value of distance data (provided, of course, that the methods we use to measure language divergence are suitably weighted and balanced).

For certain specific purposes, the tree idealization may be valid, indeed indispensable. But it is above all when it comes to representing what actually happened as a given family of languages diverged, in which configurations, and in which real-world scenarios of their speaker populations, that the tree idealization will not do. Not least when we look to phylogenetic tools, let us not allow our visions of language prehistory to become detached from the real-world forces that shape how languages diverge in the first place, as they act upon the populations that speak them. Cross-cutting relationships are nothing if not entirely normal 'facts of life' of how languages naturally diverge. Nor, in seeking to account for them, does the contrast between a branching tree with later contacts, versus a dialect continuum, lie only on the level of abstract models. Rather, viewed in terms of population prehistory, it corresponds to two quite different scenarios that the different models effectively argue for.

The utility of Germanic as a case-study is that it provides a (reasonably) known external history against which to assess our methodological approaches. On the strength of the findings here, a similar logic can now be extended to probing the unknown of how the early divergence history of Indo-European unfolded. In the full exploration in Heggarty (in preparation *a*), it transpires that even the data underlying figures 1 and 2 here suggest an early divergence along the lines of a dialect continuum. And for all the purported analytical elegance of binary branches, as a real-world demographic scenario it is this Indo-European continuum that offers the more straightforward and economical explanation. A splits-then-borrowing scenario has instead to invoke not just a complex series of

divergent migrations, but then later movements to attenuate this by bringing certain groups back into contact again. This in turn entails consequences for which of the main rival hypotheses—the migratory Kurgan 'horse culture', or the progressive demic diffusion of agriculture—best fits as the driving force that shaped the pattern of the earliest Indo-European expansion.

This research was made possible thanks to funding from the Leverhulme Trust, for the multidisciplinary *Languages and Origins* project in Cambridge (grant F/09757/A to P.H.); and from the Arts and Humanities Research Council (grant 112 229), for the project *Sound Comparisons*: *Dialect and Language Comparison and Classification by Phonetic Similarity*, based in Edinburgh (to P.H., W.M. and A.M.).

REFERENCES

Anttila, R. 1989 *Historical and comparative linguistics*, 2nd edn. Amsterdam, The Netherlands: John Benjamins.

Atkinson, Q. D. & Gray, R. D. 2006 How old is the Indo-European language family? Progress or more moths to the flame? In *Phylogenetic methods and the prehistory of languages* (eds P. Forster & C. Renfrew), pp. 91–109. Cambridge, UK: McDonald Institute for Archaeological Research.

Atkinson, Q. D., Nicholls, G., Welch, D. & Gray, R. D. 2005 From words to dates: water into wine, mathemagic or phylogenetic inference? *Trans. Philol. Soc.* **103**, 193–219. (doi:10.1111/j.1467-968X.2005.00151.x)

Bandelt, H. J. & Dress, A. W. 1992 Split decomposition: a new and useful approach to phylogenetic analysis of distance data. *Mol. Phylogenet. Evol.* **1**, 242–252. (doi:10.1016/1055-7903(92)90021-8)

Bandelt, H. J., Forster, P., Sykes, B. C. & Richards, M. B. 1995 Mitochondrial portraits of human populations. *Genetics* **141**, 743–753.

Bandelt, H. J., Forster, P. & Röhl, A. 1999 Median-joining networks for inferring intraspecific phylogenies. *Mol. Biol. Evol.* **16**, 37–48.

Bryant, D. & Moulton, V. 2004 NeighborNet: an agglomerative algorithm for the construction of phylogenetic networks. *Mol. Biol. Evol.* **21**, 255–265. See www.ab.informatik.uni-tuebingen.de/software/jsplits/. (doi:10.1093/molbev/msh018)

Bryant, D., Filimon, F. & Gray, R. D. 2005 Untangling our past: languages, trees, splits and networks. In *The evolution of cultural diversity: a phylogenetic approach* (eds R. Mace, C. Holden & S. Shennan), pp. 67–84. London, UK: UCL Press.

Dyen, I., Kruskal, J. B. & Black, P. 1992 An Indo-European classification: a lexicostatistical experiment. *Trans. Am. Phil. Soc.* **82**. See www.wordgumbo.com/ie/cmp/iedata.txt.

Forster, P. & Toth, A. 2003 Toward a phylogenetic chronology of ancient Gaulish, Celtic, and Indo-European. *Proc. Natl Acad. Sci. USA* **100**, 9079–9084. (doi:10.1073/pnas.1331158100)

Forster, P., Toth, A. & Bandelt, H.-J. 1998 Evolutionary network analysis of word lists: visualizing the relationships between alpine Romance languages. *J. Quant. Linguist.* **5**, 174–187. (doi:10.1080/09296179808590125)

Forster, P., Polzin, T. & Röhl, A. 2006 Evolution of English basic vocabulary within the network of Germanic languages. In *Phylogenetic methods and the prehistory of languages* (eds P. Forster & C. Renfrew), pp. 131–138. Cambridge, UK: McDonald Institute for Archaeological Research.

Gray, R. D. & Atkinson, Q. D. 2003 Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature* **426**, 435–439. (doi:10.1038/nature02029)

Gray, R. D., Bryant, D. & Greenhill, S. J. 2010 On the shape and fabric of human history. *Phil. Trans. R. Soc. B* **365**, 3923–3933. (doi:10.1098/rstb.2010.0162)

Heggarty, P. 2005 Enigmas en el origen de las lenguas andinas: aplicando nuevas técnicas a las incógnitas por resolver. *Rev. Andina.* **40**, 9–57.

Heggarty, P. 2006 Interdisciplinary indiscipline? Can phylogenetic methods meaningfully be applied to language data—and to dating language? In *Phylogenetic methods and the prehistory of languages* (eds P. Forster & C. Renfrew), pp. 183–194. Cambridge, UK: McDonald Institute for Archaeological Research.

Heggarty, P. In preparation *a*. Barking up the wrong Indo-European tree?

Heggarty, P. In preparation *b*. How language lineages diverge: models vs. the real world.

Heggarty, P., McMahon, A. & McMahon, R. 2005 From phonetic similarity to dialect classification: a principled approach. In *Perspectives on variation* (eds N. Delbecque, J. van der Auwera & D. Geeraerts), pp. 43–91. Amsterdam, The Netherlands: Mouton de Gruyter.

Holden, C. J. & Gray, R. D. 2006 Rapid radiation, borrowing and dialect continua in the Bantu languages. In *Phylogenetic methods and the prehistory of languages* (eds P. Forster & C. Renfrew), pp. 19–31. Cambridge, UK: McDonald Institute for Archaeological Research.

Holland, B. & Moulton, V. 2003 Consensus networks: a method for visualizing incompatibilities in collections of trees. *Algorithms in bioinformatics* (eds G. Benson & R. Page), pp. 165–176. Berlin, Germany: Springer.

Huson, D. H. & Bryant, D. 2006 Application of phylogenetic networks in evolutionary studies. *Mol. Biol. Evol.* **23**, 254–267. See www.ab.informatik.uni.tuebingen.de/software/jsplits/. (doi:10.1093/molbev/msj030)

Maguire, W., McMahon, A., Heggarty, P. & Dediu, D. 2010 The past, present and future of English dialects: quantifying convergence, divergence and dynamic equilibrium. *Lang. Variation Change* **22**, 69–104. (doi:10.1017/S0954394510000013)

McMahon, A. & McMahon, R. 2005 *Language classification by numbers*. Oxford, UK: Oxford University Press.

McMahon, A., Heggarty, P., McMahon, R. & Slaska, N. 2005 Swadesh sublists and the benefits of borrowing: an Andean case study. In *Quantitative methods in language comparison—Transactions of the Philological Society*, vol. 103 (ed. A. McMahon), pp. 147–169. Oxford, UK: Blackwell.

McMahon, A., Heggarty, P., McMahon, R. & Maguire, W. 2007 The sound patterns of Englishes: representing phonetic similarity. *Engl. Lang. Linguist.* **11**, 113–142. (doi:10.1017/S1360674306002139)

Nakhleh, L., Ringe, D. & Warnow, T. 2005 Perfect phylogenetic networks: a new methodology for reconstructing the evolutionary history of natural languages. *Language* **81**, 382–420. (doi:10.1353/lan.2005.0078)

Pagel, M. & Meade, A. 2006 Estimating rates of lexical replacement on phylogenetic trees of languages. In *Phylogenetic methods and the prehistory of languages* (eds P. Forster & C. Renfrew), pp. 173–182. Cambridge, UK: McDonald Institute for Archaeological Research.

Renfrew, C. 1989 *Archaeology and language: the puzzle of Indo-European origins*. London, UK: Penguin.

Ringe, D. A., Warnow, T. & Taylor, A. 2002 Indo-European and computational cladistics. *Trans. Philol. Soc.* **100**, 59–129. See www.cs.rice.edu/~nakhleh/CPHL. (doi:10.1111/1467-968X.00091)

Robinson, O. W. 1992 *Old English and its closest relatives*. London, UK: Routledge.