

Trees, Waves and Linkages: Models of Language Diversification

Alexandre François

▶ **To cite this version:**

Alexandre François. Trees, Waves and Linkages: Models of Language Diversification. Bowern, Claire; Evans, Bethwyn. The Routledge Handbook of Historical Linguistics, Routledge, pp.161-189, 2014, 978-0-41552-789-7. 10.4324/9781315794013.ch6 . halshs-00874726v2

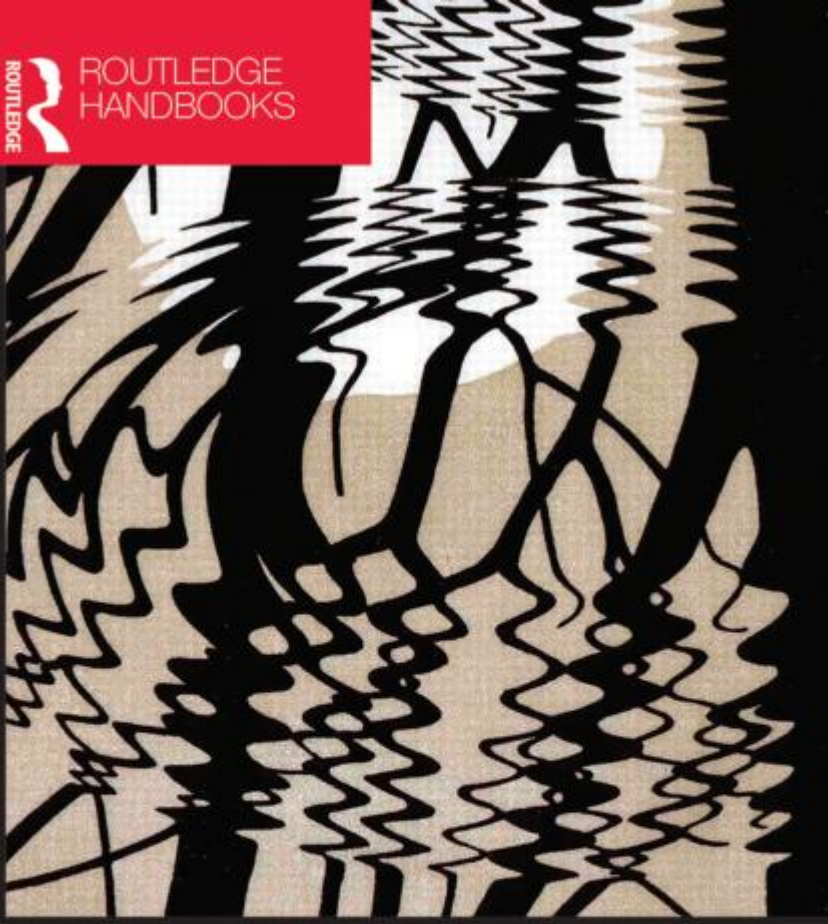
HAL Id: halshs-00874726
https://shs.hal.science/halshs-00874726v2

Submitted on 19 Mar 2015

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

ROUTLEDGE HANDBOOKS



The Routledge Handbook of Historical Linguistics

Edited by Claire Bowern and Bethwyn Evans

The Routledge Handbook of Historical Linguistics

Edited by
Claire Bowern and Bethwyn Evans

Routledge
Taylor & Francis Group
LONDON AND NEW YORK

6

Trees, waves and linkages

Models of language diversification

Alexandre François

1 On the diversification of languages

1.1 Language extinction, language emergence

The number of languages spoken on the planet has oscillated up and down throughout the history of mankind.[1] Different social factors operate in opposite ways, some resulting in the decrease of language diversity, others favouring the emergence of new languages. Thus, languages fade away and disappear when their speakers undergo some pressure towards abandoning their heritage language and replacing it in all contexts with a new language that is in some way more socially prominent (Simpson, this volume). The process of language extinction may be rapid or slow, and varies in intensity depending on historical circumstances.

While this process results in the erosion of language diversity, others bring about the opposite result: an increase in the number of spoken languages. Because no natural language appears *ex nihilo*, one has to explain how new languages emerge out of older ones. Some – such as pidgins and creoles (Romaine 1988; Siegel 2008) or mixed languages (Matras and Bakker 2003) – result historically from the encounter of two populations who were driven, under very special social conditions, to combine elements of their respective languages and create a new one. Yet this pattern, whereby a language is born of two parents, is not the typical scenario. New languages also commonly arise from the internal diversification of a single language as it evolves into separate daughter languages over time, following processes where external input does not necessarily play the central role. This phenomenon of internal diversification is the object of the present chapter.

The two tendencies outlined above – language extinction and language emergence – have always occurred in human history;[2] yet in terms of scientific knowledge, the modern scholar is faced here with a strong asymmetry. Except for the few that have left behind written materials that can be deciphered, most extinct languages of the past will forever be unknown, whether in their linguistic structures or the social causes of their demise. By contrast, linguistic diversification has brought about an observable outcome, in the form of attested languages. The latter can be analysed and compared in a historical perspective, thereby bringing invaluable insights into their linguistic and social development. This asymmetry in the availability of data explains why the process of *language diversification* plays such a

central role in the discipline of historical linguistics. The aim of the present chapter is to understand how this process of diversification takes place in languages, and what model can best account for the empirically observed patterns of language relations.

1.2 Trees vs. waves: two models of language diversification

Our point of departure is the observation that several modern languages can historically stem from the internal diversification of what was once a single language, with no need to resort primarily to external factors such as contact or language admixture. The internal diversity among modern Romance languages, for example, can largely be explained by a process of internal fragmentation, taking a relatively homogeneous variety of spoken Latin as a starting point. While contact-related factors – substrate, superstrate and adstrate influences involving non-Romance languages – did play their part, a large proportion of the history of Romance can be reconstructed as internal diversification affecting inherited linguistic material.

For most language families, unlike in Romance, the ancestral language is not attested but merely hypothetical; the reconstruction of historical scenarios leading to modern languages is then the object of logical analysis and the weighing of competing hypotheses, based on a systematic comparison of the attested languages. This procedure, known as the Comparative Method (see chapters by Weiss and Hale in this volume), was initially developed by the German Neogrammarians in the second half of the nineteenth century, and constitutes, to this day, the most successful approach in reconstructing the history of language families.

The Comparative Method has tended to be closely associated with a particular model of diversification: the *Stammbaum*, or family tree. Ever since this model was first proposed by August Schleicher in his 1853 article *Die ersten Spaltungen des indogermanischen Urvolkes*, its association with the Comparative Method has been taken for granted (e.g. Bloomfield 1933: 311; Campbell 2004: 165; etc.); yet I will claim here that the two lines of thinking ought to be dissociated. While the Comparative Method is without a doubt the most solid approach to the reconstruction of language histories, I will argue that the Tree Model presupposes a flawed understanding of language diversification processes. In a nutshell, cladistic (tree-based) representations are entirely based on the fiction that the main reason why new languages emerge is the abrupt division of a language community into separate social groups. Trees fail to capture the very common situation in which linguistic diversification results from the fragmentation of a language into a network of dialects which remained in contact with each other for an extended period of time (Bloomfield 1933; Croft 2000; Garrett 2006; Heggarty, Maguire and McMahon 2010; Drinka 2013), creating what Ross (1988, 1997) calls a 'linkage' (see section 3.3).

The present chapter will discuss the strengths and weaknesses of cladistic representations for modelling processes of language diversification, and examine alternative approaches for capturing the genealogy[3] of languages. In section 2, I will first summarise the way in which linguistic trees are typically understood, before examining their underlying assumptions. Section 3 will examine the processes that underlie genealogical relations between languages, and explain why the Tree Model is most often unsuited for representing them. While the Comparative Method must be preserved for its invaluable scientific power, a rigorous application of its principles in situations of linkage in fact disproves the Tree Model, and favours the Wave Model (section 3.2) as a more accurate description of the genealogy of languages.

Non-cladistic models are needed to represent language relationships, in ways that take into account the common case of linkages and intersecting subgroups. Among existing models, Section 4 will focus on an approach that combines the precision of the Comparative Method with the realism of the Wave Model. This method, labelled *Historical Glottometry* (Kalyan and François forthcoming), identifies genealogical subgroups in a linkage situation, and assesses their relative strengths based on the distribution of innovations among modern languages. Provided it is applied with the rigour inherent to the Comparative Method, Historical Glottometry should help unravel the genealogical structures of the world's language families, by acknowledging the role played by linguistic convergence and diffusion in the historical processes of language diversification.

2 Understanding the tree model

2.1 Reading and drawing language trees

I first propose to examine how language trees are classically understood. Let there be five modern languages, labelled K, L, M, N, O. These languages are believed to be genealogically related if they comply with a number of conditions (Campbell and Poser 2008: 162ff): in particular, a sizeable number of demonstrably cognate items in their morphology and basic vocabulary, displaying regular sound correspondences in ways that cannot be reasonably assigned to chance or borrowing (Weiss, this volume).

To say that K, L, M, N, O are genealogically related entails that they ultimately descend from a common ancestor – a 'proto-language', which in this case can be called Proto-KLMNO. This point could be shown using Figure 6.1, a 'rake-like' or 'fan-like' representation: this shows each language as an independent descendant of the proto-language, with no claim about the family's internal structure. Such a 'flat' tree may sometimes correspond to an actual historical situation, as when an ancestral society swiftly broke up into a number of separate sub-communities, quickly followed by a loss of mutual social contact; according to Pawley (1999), this scenario may indeed have characterised the break-up of Proto-Oceanic into lower-level subgroups. In other cases, a representation like Figure 6.1 simply reflects a linguist's agnostic view of a family's internal structure, for instance due to lack of sufficient data. What historical linguists typically hope to achieve with a tree is to identify a number of internal subgroups within the family, into which languages with more recent shared ancestors can be grouped together. Figure 6.2 illustrates the sort of ideal tree aimed at by subgrouping studies.

Such a tree captures a set of claims about the internal structure of a language family. Here, a claim is made that languages K and L 'subgroup' together, by contrast with M, N and O which form their own subgroup MNO; within the latter, a claim is made that N and O form a subgroup of their own apart from M. Following a nested pattern, the language N is said to belong to the NO subgroup, which in turn forms a 'branch' of the larger subgroup MNO. Even though such claims about the internal structure of a family could be formulated, in principle, in purely taxonomic terms with no reference to time, it is common practice to interpret such cladistic representations of language families in historical terms. A common assumption is that the sequence of nodes in a tree, from top to bottom, mirrors the actual chronological order of historical events. Another frequent, and somewhat simplistic, conception (as underlined by Pulgram 1961) is that each node in the tree corresponds to an individual language community, so that a split in a tree can essentially be equated with the division of an earlier unified community into separate social groups.

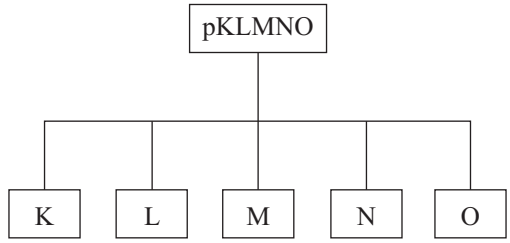


Figure 6.1 An unordered genealogical tree

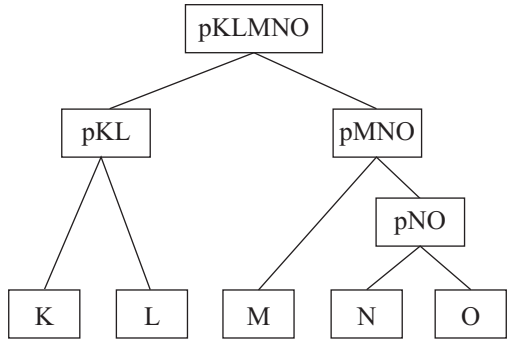


Figure 6.2 A genealogical tree indicating internal subgrouping

Thus, to say that M, N and O subgroup together as opposed to other languages of their family, amounts to claiming that they all descend from an intermediate proto-language – call it Proto-MNO – that was once spoken by a single social community, after the break-up of the earlier language Proto-KLMNO. According to Figure 6.2, this language Proto-MNO must have developed more or less separately from Proto-KL, the shared ancestor of modern languages K and L. This point is established through the identification of a number of linguistic innovations of various sorts (phonological, grammatical, lexical, etc.) which are jointly reflected by modern languages M, N and O, but not by other languages of the family. If these three languages share together certain linguistic properties that were not inherited from their ultimate ancestor, it is assumed – provided one can rule out chance similarity or parallel innovation – that they must have acquired these properties at a certain point in time, when their speakers still spoke (mutually intelligible variants of) a single language. The idea is that, instead of positing the same change in three languages (M, N, O) independently, it is more parsimonious – following Occam's razor – to propose that it took place just once in a single language (Proto-MNO) and then was simply inherited by its descendants. By contrast, the fact that K and L do not reflect those innovations suggests that their ancestors did not participate in that Proto-MNO speech community. This scenario is visually summarised by the existence of the 'MNO' node in Figure 6.2.

Following a principle first formulated by Leskien (1876), the Comparative Method establishes the existence of every intermediate node in a family tree based on the principle of *exclusively shared innovations*, i.e. by identifying those linguistic changes that are shared by all of its modern descendants, and only by them – what phylogeneticists call *synapomorphies* (Page and Holmes 2009). These innovations are thought to have been introduced historically during the lifetime of the intermediate proto-language – after the split from a higher node,

and before the new split into lower nodes. The reasoning is recursive: Figure 6.2 also represents the claim that the ancestors of modern speakers of M, after undergoing developments that are also reflected in N and O, at some point in time started developing independently; by contrast, the remaining ancestors of N and O kept sharing innovations for some time, until they too eventually separated.

In sum, the history of the family illustrated in Figure 6.2 would be summarised by stating that what used to be a single language (pKLMNO) first split into two separate languages (pKL and pMNO), which in turn were to split again. This series of recursive splits and the resulting divergence is one possible way to understand the process of language diversification, and the emergence of new languages.

2.2 The tree, a model based exclusively on separation

In the classical understanding of family trees, each node is thus supposed to correspond to a specific social community that developed separately from other nodes (Fox 1995: 123). The sort of separation referred to here is typically understood as an actual event of social split such as migration, whereby a previously unified society broke up into two separate communities with loss of contact. Other cases are possible, such as social isolation due to the intrusion of other languages; or the *in situ* break-up of earlier networks of communication, as communities stayed in place yet decreased their mutual contact as they began – for whatever reason – to isolate themselves from each other.[4]

In order to yield a robust tree-like structure like the one in Figure 6.2 with intermediate nodes (as opposed to the flat structure of Figure 6.1), the process of social split must be repeated recursively across the centuries; each event of separation must have been followed by a period of stability – at least a few generations – during which innovations had the time to form and settle within the new community (Pawley and Ross 1995), before another split took place again.

This focus on divergence is both a strength and a weakness of the Tree Model. A strength, because it means that trees can help reconstruct events of social disruption when they indeed took place, and can represent them using a visually straightforward diagram. But it is also a weakness, because it distorts the reality of language diversification by shoehorning it into a one-size-fits-all, simplistic model which forces us to reconstruct events of social separation even when they never really happened, at the expense of all other possible scenarios.

Let us imagine, for the sake of discussion, that there existed a language family in the world whose development did indeed take the form of social splits, repeated over and over through the centuries of its history: such a hypothetical language family could indeed be portrayed accurately by a tree such as Figure 6.2 above. In reality, no population in the world can reasonably have its history reduced to just a series of social splits with loss of contact – the scenario favoured by the Tree Model. While some families did go through such events several times in their history, in the form of successive bouts of migration or similar disruptions, these events of split, correlated with neat patterns of linguistic divergence, are always interspersed with other forms of social interaction whose linguistic impact – as we'll see below – is not compatible with a tree representation.

2.3 Dealing with problems in a tree structure

In the interest of the forthcoming discussion, it is important to highlight the fact that, under the Tree Model, a given language may belong to only one higher-level subgroup at a time. If

M is a member of the MNO subgroup, then it cannot also be a member of a KLM subgroup at the same time: subgroups are mutually exclusive, and never intersect. This seems a sensible idea if the splits in the tree are meant to represent physical separation with no return: if the communities of pKL and pMNO were indeed separated with complete loss of contact, then it is difficult to imagine how some modern descendants of pMNO, but not others, could share anything with pKL. This principle of separate development is central to the whole logic of subgrouping studies under a cladistic approach, and has important consequences.

Let's assume that the tree in Figure 6.2 rests on sufficiently solid evidence to be deemed valid. Then let's imagine that a linguistic property is found to be shared by languages L and M, and only these two languages. This will be a problem under the Tree Model, one that will require specific hypotheses in order to account for this shared property, and still save the tree structure. For example, the shared property may be proposed to be in fact a case of *shared retention* (also known as *symplesiomorphy* in phylogenetics) from the Proto-KLMNO ancestor, a property lost by other languages (K, NO): in this case, the property would not indicate any significant genealogical link between L and M – other than their remote relatedness. Alternatively, one could argue that the property is indeed innovative, yet happened independently in L and M, whether by drift or parallel innovation (*homoplasy*).

Finally, a third hypothesis would be that the property was innovated internally in only one language, say L, and then was borrowed by the other language M via *contact* between L and M, once they had already been formed as separate languages. Even though contact between languages – also known as 'horizontal transmission' or 'areal diffusion' – is known to be a powerful force of language change (Lucas, this volume), it is not meant to be represented on a tree. Contact-induced change, which can take place between any two languages regardless of their relatedness, is generally considered to be a separate phenomenon from the sort of 'internal change' that underlies genealogical relations. The argument is that, for a property to be borrowed between two separate languages L and M, the two languages need to already exist independently; strictly speaking, the study of their genealogy is interested in how these languages came into existence, not in what happened to them later. Thus, the many words borrowed by English from Scandinavian languages during the Viking invasions, or later from French, are not considered to form part of its genealogical makeup: the English language had by that time already acquired independent existence, as it were, as a member of the Anglo-Frisian branch of the West Germanic subgroup. Following this principle, in a tree such as Figure 6.2, a property borrowed by M from L after their separation would not be considered evidence for a genealogical subgroup LM; it would be described as an effect of contact, and understood as irrelevant for subgrouping purposes.

Several authors have expressed frustration at the Tree Model, saying that trees exclusively represent language divergence, and fail to take into account contact-induced change, or *convergence*, when reconstructing language history (e.g. Fox 1995: 124; Dixon 1997; Aikhenvald and Dixon 2001; Bossong 2009; Drinka 2013). They argue that loanwords, borrowed structures and other facts of cross-linguistic diffusion form part of the linguistic history of languages as much as the material directly inherited. While the latter point is undoubtedly true, proponents of the Tree Model reply to this objection by acknowledging that trees are only intended to capture a portion of the history of languages, namely their *genealogy* strictly speaking, and nothing more. As for other facts of language development – notably the effects of contact – they are, or at least should be, treated by other models (Campbell and Poser 2008: 327). This is a valid point, which bears keeping in mind every time family trees are cited: language genealogy only forms a portion of the historical picture, and trees should not be assigned more explanatory value than they actually have.

In the following sections, the argument I will put forward against the Tree Model is reminiscent of the objection just mentioned, yet distinct from it. Let us grant that contact *between separate languages* (e.g. Old English and Old French) does not form part of their genealogical makeup, and that the model we want to design is meant to focus on the latter. My main proposal will be that trees not only omit representing language contact (which is fair enough, if it is not their objective to do so) but also, more problematically, that they even fail to accurately represent language genealogy. My argument will also be based on the problem of horizontal diffusion; yet instead of concerning facts of *cross-linguistic diffusion* (contact between already separated languages), my central problem will be processes of *language-internal diffusion* – i.e. the diffusion of innovations across mutually intelligible idiolects in a single language community.[5]

For example, the whole reasoning above about a property shared between L and M would have to be quite different if the KL and MNO clusters were never in fact physically separated, but were simply sets of dialects within a larger KLMNO group of mutually intelligible varieties still in constant contact. While it may be the case that dialects KL have shared together one set of innovations and MNO another one, it is perfectly plausible that dialects L and M could *also* undergo their own set of shared innovations, during the same historical period. This is how the process of language-internal diffusion, the ultimate source of genealogical relations in languages (section 3.1), can give birth to subgroups that crosscut each other: KL; LM; MNO… Such a dialect-chain situation, and more generally dialect continua and linkages (section 3.3), form the Achilles' heel of the Tree Model, and are best described using a non-cladistic approach (Gray, Bryant and Greenhill 2010: 3229). This issue is the focus of the next section.

3 The waves of diffusion at the source of language genealogy

3.1 Theoretical principles: genealogy reflects diffusion

Recent progress made on the sociolinguistic underpinnings of language change provides an opportunity to rethink the process of linguistic diversification, and to redefine what we mean by 'genealogical' or 'genetic' relations in languages. In particular, one assumption held by the founders of the Tree Model was that the normal locus of linguistic innovations is a 'language' or a 'proto-language', understood as a monolithic unit that could be represented as a simple node in a tree. Thus for modern languages M, N and O to share the same innovation *i* would be interpreted as evidence that these necessarily descended from a single *language* (labelled Proto-MNO). Positing such a node in the tree makes it then possible to state that the innovation *i* happened just 'once' in that single language – with the assumption that this would be more parsimonious than positing parallel innovation or late contact between three separate languages M, N, O (section 2.1). The whole design of the family tree rests on this fiction that a 'language' unproblematically forms an atomic unit, and that innovations just 'happen' in them.

This simplistic view was challenged as early as the end of the nineteenth century by the work of dialectologists (Gilliéron 1880; Wenker 1881), who showed that a given language typically consists of a network of dialects that can show a great deal of diversity. Language properties were found to be distributed in space following complex patterns, described visually using *isoglosses*. Far from always coinciding neatly, the default situation for these isoglosses is to target different segments of the social network, and thus intersect (cf. Trudgill 1986; Fox 1995: 129; Chambers and Trudgill 1998). These views from dialectology were enriched by sociolinguistic studies, which observed how individual instances of language

change are reproduced and diffused by speakers in their daily communication (see Labov 1963, 1994, 2001, 2007; Milroy 1987; Milroy and Milroy 1985). These works emphasised not only the complex geographical distribution of properties, but the intricate patterns whereby tokens of innovative features are statistically distributed in the speech of individuals, depending on a variety of social factors.

When approaching language change, the proper operational unit of observation is not the language or the dialect, but the *idiolect*, i.e. the linguistic competence of an individual speaker at a certain point in time.[6] As for dialects and languages, they form more or less homogeneous systems shared by a network of *mutually intelligible idiolects*. When historical linguists identify a change that happened 'once' in a 'language', they really encapsulate a long process of diffusion that took place across large networks of idiolects, sometimes spanning across several generations.

Indeed, linguistic innovations first emerge in the speech of certain individuals, in the form of novel ways of speaking – whether phonetic, lexical, phraseological, etc. If that innovation presents some sort of appeal to the hearer as a way to potentially increase their communicative goals, they may adopt it into their own speech, and start replicating it in new situations. If carried out repeatedly and extensively across a social network, this process of imitation or 'accommodation' (Street and Giles 1982; Trudgill 1986; Giles and Ogay 2007) results in the spread of a new speech habit from one person to the other, across idiolects – a phenomenon which has been labelled *propagation* (Croft 2000) or linguistic *epidemiology* (Enfield 2003, 2008). After a period of competition with the previous norm, the innovation may become statistically dominant, and settle in the speech habits of a whole social group. If it does, then it becomes a property of an entire 'communalect' (i.e. sociolect, dialect or language). From that point onwards, the linguistic feature will be transmitted to descendant generations of learners, just as much as the rest of the inherited system.

This language-internal diffusion of innovations gives rise to the genealogical relations among languages which subgrouping studies precisely seek to unravel. Such a process is not fundamentally different from what is involved in language contact: both forms of diffusion involve the progression of a new linguistic behaviour across a social network of individual speakers – a phenomenon which is not reducible to a single event. The main distinction is that *contact* is normally a process of diffusion observed across separate languages, whereas *language-internal diffusion* involves mutually intelligible idiolects, which together may be taken to form a single (more or less homogeneous) language community.[7]

An innovation diffusing through a community may eventually (sometimes after several generations) settle into the mainstream usage of an entire network of mutually-intelligible idiolects, and thus become a feature of 'the language'. When this happens, the change may be captured using a synthetic formula of the type $x > y$; it may even be understood, in retrospect, as though it were a single event that took place 'once' in that 'language'. However, the patterns of propagation are often more complex. Specifically, the language-internal diffusion of innovations does not have to target an entire language community, and commonly settles down to just a cluster of dialects, so that successive innovations target different segments of the network. In this case, the intricate patterns resulting from language-internal diffusion cannot be captured by a tree, and need to be accounted for by a different model.

3.2 The Wave Model

Just such a line of theoretical reasoning underlies the 'Wave Model', or *Wellentheorie*, which Hugo Schuchardt and Johannes Schmidt proposed in the early 1870s (Schmidt 1872), as an

alternative to August Schleicher's Tree Model (*Stammbaumtheorie*). These authors occasionally conceived their Wave Model as a challenge not only to the Tree Model, but to the Comparative Method as a whole: Schuchardt, for example, linked it with a general disbelief in the Neogrammarians' views on the regularity of sound change (Schuchardt 1885). Such an extreme stance is however not essential to the Wave Model, and unduly throws the baby (the Comparative Method) out with the bathwater (the Tree Model). A synthesis should be possible, which preserves the principle of regularity and other useful tenets of the Comparative Method, yet replaces the simplistic tree representations with a wave-inspired approach.

Under the Wave Model, each instance of language change arises somewhere within the network, and from there diffuses to adjacent speaker groups. The propagation of the change can thus be compared to a 'wave' which expands away from its centre as the new feature is adopted across a broader territory. These waves are independent of each other, and are not necessarily nested. As Bloomfield (1933: 317) puts it, "[d]ifferent linguistic changes may spread, like waves, over a speech-area, and each change may be carried out over a part of the area that does not coincide with the part covered by an earlier change." Likewise, an innovation targeting a small cluster of dialects can be followed by a later one targeting a larger group.[8] Both these patterns are incompatible with a tree.

I will illustrate this point first with a general model, before mentioning actual examples. Each event of language change defines its own isogloss, i.e. a (typically) geographically contiguous zone, representable on a map, within which the innovation diffused across idiolects and settled. In a linguistic continuum characterised by mutual intelligibility across adjacent dialects, the normal situation is for these isoglosses to intersect constantly, rather than be nested. For instance, let there be eight close dialects labelled A to H, and six innovations numbered #1 to #6, such that innovation #1 arose in dialect D and spread to adjacent dialects until it covered the zone CDE; #2 encompassed AB; #3 spanned over CDEF; #4 over FG; #5 over EF, and #6 over EFGH (Figure 6.3).

The first innovations which targeted, say, the dialects CDE, were not radical enough to prevent mutual intelligibility with the other dialects: in the absence of a physical boundary between them, nothing then prevented the next innovation from targeting a cluster EF, then FG, etc. In this model, every innovation constitutes an instance of linguistic *convergence* – for the dialects that participate together in that innovation, e.g. E and F in #5 – as much as it is a case of linguistic *divergence* – for the dialects that become differentiated as a result of the change, e.g. E and D in #5 (cf. François 2011a: 231).

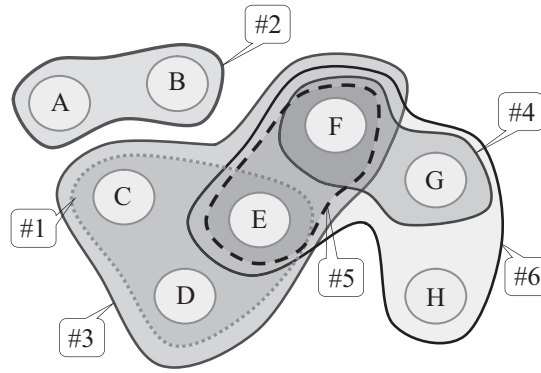


Figure 6.3 Intersecting isoglosses in a dialect continuum or a linkage

Over time, the layered innovations leave their footprint in each local dialect. Consider a pair of dialects, for example F and G. While the changes they share together (#4, #6) have increased their similarity in certain aspects of their systems, those which have affected only one of them (either alone, or together with other neighbouring dialects – e.g. #3, #5) have increased their difference. Should many more cross-cutting innovations (or 'non-shared innovations') accumulate over generations, what started as mutually intelligible dialects F and G will become opaque to each other, and eventually become distinct languages. Unless later processes of dialect levelling (or *koineisation*) take place,[9] each member of the network will inherit in its local system the innovations it has participated in, and these will be transmitted to its descendants. In this regard, all the innovations mentioned here, and represented in Figure 6.3, define the genealogical structure of the family.

As these dialects increase their differences and lose mutual intelligibility, the end result is a rise in the number of distinct languages. Yet crucially, whereas the Tree Model assigns linguistic diversification to social splits with loss of contact (section 2.2), the Wave Model is compatible with scenarios where communities remain in contact. In fact, it treats linguistic contact – in the form of multiple, criss-crossing events of diffusion across mutually intelligible dialects – as the very key to understanding patterns of language diversification. This is a radical shift in perspective.

An important implication of the Wave Model is that a given language can perfectly well belong to several partially overlapping subgroups. A genealogical subgroup is here defined as a group of languages whose ancestors participated together in the diffusion of one or several linguistic innovations, at a time when they were mutually intelligible. Crucially, nothing in this definition entails that subgroups should be discrete or nested, and indeed my claim is that genealogical subgroups can perfectly intersect, and commonly do. Thus in Figure 6.3, it is legitimate to say that E belongs simultaneously to the subgroups CDE, EF, CDEF, and EFGH – a situation which no orthodox family tree would ever be able to represent (section 2.3).

3.3 From dialect continua to linkages

The issue of isogloss intersection has long been central to dialect geography (see Bloomfield 1933: 321). It thus comes as no surprise that dialectologists, who observe the fine-grained distribution of linguistic features in space, tend to favour the Wave Model – or some model derived from it – over cladistic representations. The networks of Italian, Dutch or Arabic dialects, to take just a few examples, could never be modelled by any tree.

One could propose that the two models are complementary, in the sense that trees would be well-designed to represent the genealogical relations between separate *languages*; whereas waves would only be concerned with the complex relations between *dialects* within the boundaries of each language. The two models would then both be useful, but at different grains of observation. I think this view is wrong, for one important reason: namely, that many language (sub)families – as we will see below – have in fact arisen from the diversification of former dialect continua. To the extent that earlier local innovations are faithfully transmitted across generations,[10] the resulting languages normally keep the traces of their entangled isoglosses. If trees fail to represent genealogical relations between dialects, then they must also fail to capture the relations between the languages that descend from them.

This important point has been made by Malcolm Ross, around the concept of *linkage* (Ross 1988, 1996, 1997, 2001). Ross (1988: 8) defines a linkage as "a group of communalects which have arisen by dialect differentiation," where 'communalect' is a generic term which

may refer to modern dialects or languages. When a dialect continuum – typically structured along the lines of Figure 6.3 above – evolves in such a way that its members lose mutual intelligibility, it becomes a linkage. A linkage thus consists of separate modern languages which are all related and linked together by intersecting layers of innovations; it is a language family whose internal genealogy cannot be represented by any tree.

While Ross initially developed this concept for the historical reconstruction of Western Oceanic languages, it clearly has wider theoretical significance. Many language families or subfamilies have been shown to be linkages – whether the authors have used that term or not. The Oceanic languages of Fiji (Geraghty 1983), Polynesian languages (Gray, Bryant and Greenhill 2010), the Indo-Aryan languages of the Kamta region of India (Toulmin 2006, 2009), the Karnic subgroup of Pama-Nyungan (Bowern 2006), northern Athabaskan languages (Krauss and Golla 1981; Holton 2011), some parts of the Semitic family (Huehnergard and Rubin 2011), Sinitic languages (Hashimoto 1992; Chappell 2001), Western Romance (Penny 2000: 9–74; Ernst *et al.* 2009), Germanic (Ramat 1998), and even Indo-European as a whole (Bloomfield 1933: 316; Anttila 1989: 305; Garrett 2006; Drinka 2013): these are all examples, among many others throughout the world, of language families which have been shown to result from a long history of layered innovations with entangled patterns of distribution, akin to Figure 6.3 above; none of them could be accurately represented by a tree. Section 4.3 below will briefly examine a particular linkage from northern Vanuatu, and propose a way to model such linkages.

3.4 The tree, a special case of a linkage

Based on empirical observation of the world's language families – as illustrated by the scholarly works cited above, and many others – it thus seems that genuinely 'tree-like' families are much rarer than is usually acknowledged. This is so true, that one may question the usefulness of the Tree Model as a suitable approach for representing language genealogy altogether.

One might perhaps propose to salvage the Tree Model as a useful approximation, at least for those (sub)families which are mostly compatible with it. This would go along with the conventional wisdom that the Tree and the Wave models complement each other, and should both be preserved (Hock 1991: 454; Rankin 2003: 186; Labov 2007, etc.). However, this conclusion does not appear necessary, because a tree-like structure is nothing more than a special case of a linkage – an exceptional case in which isoglosses just happen to be nested, and temporally ordered from broadest to narrowest.

And indeed, an important strength of the wave approach is its ability to represent not only cases of cross-cutting isoglosses, but also so-called 'tree-like' situations when this is in fact appropriate. Imagine that, in Figure 6.3 above, the members of the AB subgroup were found to share no innovation at all with the other members of the family: this is shown by the absence of any isogloss involving A, B or AB together with other languages. Such an observation may reflect the fact that the ancestors of modern speakers of A and B isolated themselves from the rest of their family, whether due to social attitudes or to physical constraints – including migration with loss of contact. What would then obtain is precisely the sort of neat social split that is represented all the time by trees.

Would such social-split signals justify preserving the Tree Model? Not necessarily, for two reasons. First, even if the existence of a separate AB cluster could be represented visually by a 'branch' linking Proto-ABCDEFGH to Proto-AB, the entangled isoglosses among CDEFGH would still be incompatible with a tree, and would need to be represented by waves

anyway. All in all, a wave diagram such as Figure 6.3 is both necessary and sufficient to display the splits in question, and a tree would add nothing more.

The second argument is of a more epistemological nature, and still favours the Wave Model even in situations of neat social split. Under the Tree Model, splits are assumed to be the only force underlying the formation of subgroups; this constitutes an aprioristic axiom for the whole model to hold together. By contrast, under a Wave approach, the identification of such splits is an empirical – and falsifiable – result of observation. In terms of historical reconstruction, this is an invaluable advantage of the latter method. In other words, Waves are not only better designed than Trees for tackling entangled situations of dialect continua and linkages; they even do better at detecting cases of neat split, which the cladistic model merely takes for granted.

3.5 Synthesis: two competing models of language diversification

In sum, trees and waves constitute two competing attempts at representing the same thing, namely historical events of early language-internal 'horizontal' diffusion, apprehended through the traces they left in modern languages, via later 'vertical' transmission. Both approaches are equally concerned with diffusion (*shared innovations*) and with transmission (*shared inheritance*). They target the very same domain (*pace* Campbell and Poser 2008: 399), and it is indeed *genealogical* relations that I claim are better represented by waves than by trees.

Insofar as the Wave Model is agnostic as to whether genealogical subgroups should be expected to be nested or to intersect, it constitutes a more encompassing and flexible view of language diversification than the Tree Model; the latter approach entails a number of assumptions and simplifications which are not warranted by what we now know of the actual life of languages. In lieu of trees, historical linguists should use the Wave Model – or some approach derived from it – to achieve a more exact and realistic representation of the genealogical structure of the world's language families.

4 A model for capturing intersecting subgroups

What we need then is a method for identifying and representing genealogical subgroups when they intersect. Among several existing proposals for non-cladistic models (section 4.1), this final section will focus on one possible way of formalising the Wave Model: Historical Glottometry.

4.1 Alternative approaches to genealogy

One possible reason why trees have remained pervasive in historical linguistics, despite their long-recognised flaws, is a relatively trivial one: namely, that they offer a visually elegant and easy-to-read graphical representation of a simple scenario. For the more realistic wave-based approach to ever be fully rehabilitated, then, it is necessary to design a model that readily lends itself to readability and straightforward interpretation, without compromising empirical accuracy.

Various proposals have been made to address the flaws inherent in Schleicher's *Stammbaum*. In recent years, several phylogenetic studies have tackled the issue of weakly defined subgroups, by using Bayesian maximum-likelihood methods to assess the degree of support for each subgroup in a tree (e.g. Dunn *et al.* 2008; Greenhill and Gray 2009; Greenhill,

Drummond and Gray 2010; Gray, Bryant and Greenhill 2010; Bowern and Atkinson 2012; see Dunn, this volume). These welcome methods avoid a simplistic reading of family trees, and provide empirical ways to gauge the validity of tree-based genealogical hypotheses. Yet these are still cladistic approaches: faced with a linkage-type family, they can quantify the degree to which the family is '(non-)tree-like'; but they do not provide a convincing alternative representation of their own, which could be used to identify the precise patterns of intersection between genealogical subgroups.[11]

Other proposals have been more clearly inspired by wave or network-based representations: Southworth's (1964) 'tree-envelopes'; Anttila's (1989: 305) isogloss map of major Indo-European subgroups; Hock's (1991: 455) 'truncated octopus-like tree'; van Driem's (2001: 403) 'fallen leaves'; Forster, Toth and Bandelt's (1998: 185) 'evolutionary network'; Ross' (1997: 223, 234) social network representations of language change, etc. Each proposal contributes to the search for a representation of language genealogies that is free from the constraining assumptions of the Tree Model. However, most of them are intuitive and programmatic, and have not been applied to detailed empirical data from specific language families.

An exception must be made for the network representations in Forster *et al.* (1998) mentioned above, as well as for NeighborNets, which have recently gained wide acceptance (Bryant, Filimon and Gray 2005; Heggarty *et al.* 2010). These networks are capable of displaying pairwise distances among taxa in the form of intersecting groupings. Free from the assumptions of trees, NeighborNets make it possible to visually capture the tangled webs typical of most language families, particularly linkages. An example of such a NeighborNet will be presented, and criticised, in section 4.3.4.

Among various other approaches to representing language diversity, one should also mention *dialectometry* (Séguy 1973; Guarisma and Möhlig 1986; Goebl 2006; Nerbonne 2010; Szmrecsanyi 2011). This family of methods is used to visualise pairwise linguistic distances across a dialect continuum, calculated on the basis of large amounts of data; its results typically take the form of choropleth maps. Inspiring though it is, this approach does not attempt to tackle language history *per se*: following accepted practice among dialectologists, its assessment of linguistic distance is based merely on the comparison of synchronic features, without distinguishing shared inheritance from *shared innovations*.

4.2 Crossing the Comparative Method with the Wave Model: Historical Glottometry

In the final part of this chapter, I propose a synthesis of the theoretical principles discussed earlier, and outline a new model I call *Historical Glottometry*. This method aims at combining the precision and realism of dialectological approaches (especially dialectometry, from which its name is inspired) with the reasoning power of the Comparative Method. The objective of Historical Glottometry is to identify genealogical subgroups in a language family, and measure their relative strengths so as to assess their historical patterns of distribution across social networks. Stronger linguistic ties can then be taken as indicators of stronger bonds among past societies – precisely the sort of invaluable insight language historians hope to achieve.

Because the model here defined is meant to capture the unfolding of historical events which underlie language diversification, the focus of our attention needs to be not just on the synchronic properties of languages, but on those properties that are thought to reflect *shared innovations* – in accordance with Leskien's principle (see section 2.1). This key principle of the Comparative Method can perfectly well be applied to a wave-based or network-based

approach: this is how, for example, Figure 6.3 above should be interpreted, with each isogloss corresponding to one or more shared innovations.[12]

The tools for distinguishing innovations from retentions are also those of the Comparative Method, and will be illustrated in section 4.3.2 below; they include the principle of regularity in sound change, hypotheses on the direction of change and on relative chronology, among other principles. In this respect, the procedure is identical to the one used to identify innovations in a cladistic approach. Likewise, the Comparative Method has often proven capable of distinguishing, for example based on the observation of regular and irregular sound changes, which properties were inherited or acquired early in a dialect continuum, and which ones were acquired later by contact across already separated languages (e.g. Biggs 1965 for Rotuman). Such tools are powerful for isolating the relevant genealogical data that will feed into our historical argumentation.

Once a number of innovations have been identified, one can observe which languages have evolved together over time. Whenever a group of languages share together one or several innovations that can reasonably be assigned to processes of language-internal diffusion, they define a (more or less well-supported) genealogical subgroup (section 3.2). For each subgroup, its number ε of 'exclusively shared innovations' is a measure of how frequently its members tended to imitate each other's speech (as opposed to that of non-members), and provides a first approximation to the strength of their social bonds. For example, in Figure 6.3 above, suppose that languages EF shared 32 innovations, and FG just 8: such a linguistic measure would show that the community F had much stronger social bonds with E than it had with G.

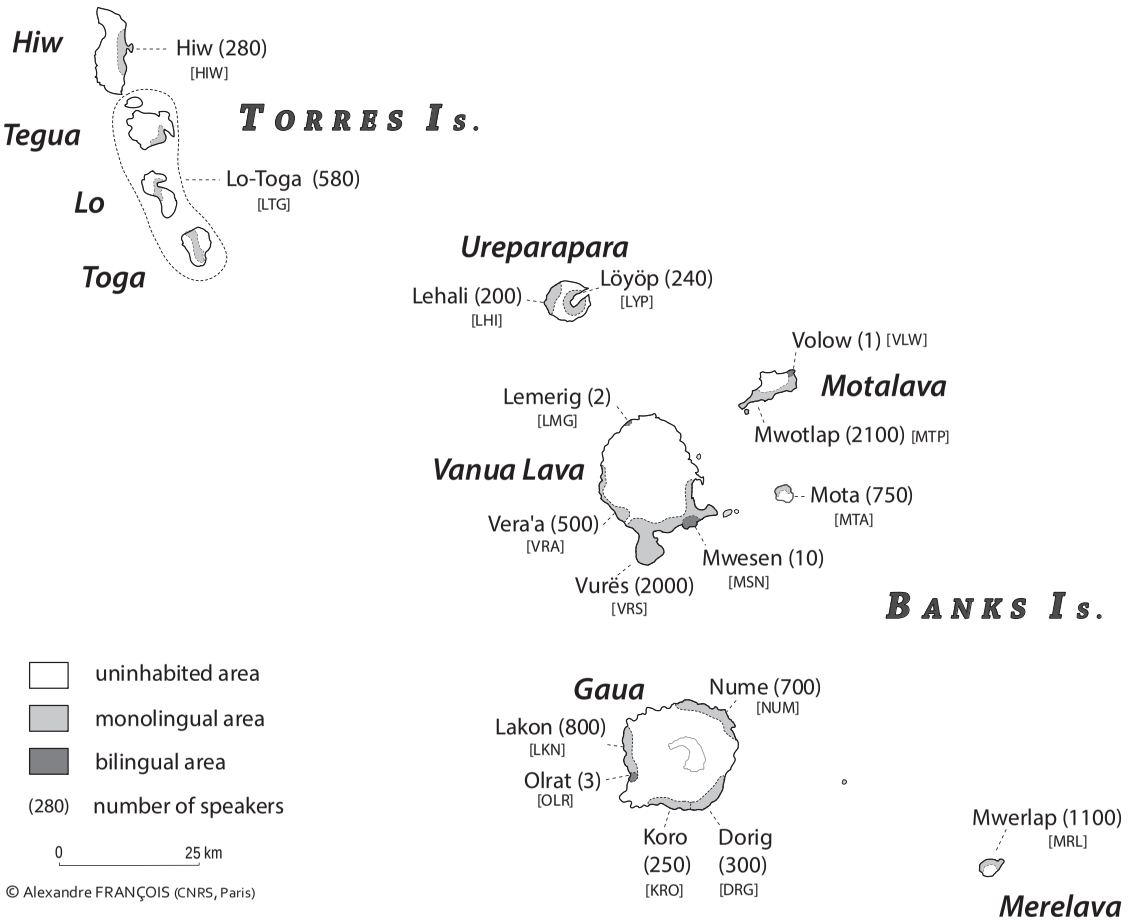
Historical Glottometry (as described in greater detail in Kalyan and François, forthcoming) provides still more precise tools to measure the relative strengths of subgroups in a linkage situation – in particular, calculations of *cohesiveness* and *subgroupiness*. These will be briefly presented below, based on actual data taken from the languages of northern Vanuatu.

4.3 A glottometric study of the northern Vanuatu linkage

4.3.1 The northern Vanuatu linkage

Vanuatu, an archipelago of island Melanesia in the South Pacific, is home to 113 indigenous languages. They all descend from Proto-Oceanic (POc), a language that was spoken about 3,200 years ago by those who first settled most of the islands of the Pacific (Pawley 1999). Apart from three Polynesian languages which arrived in Vanuatu in the last millennium, the remaining 110 languages form a linkage (Tryon 1996; Lynch 2000: 181; François 2011b): their modern diversity results from three millennia of *in situ* fragmentation, with no notable external input. This diversification was brought about by the accumulation of partially overlapping isoglosses among what started as a vast dialect network, and progressively became the linguistic mosaic we know today.

Among these 110 languages, 17 are spoken in the Torres and Banks Islands in the north of the country, by a population which has always sustained traditions of interisland marriage and social contact of various kinds (François 2011a, 2012). The names of these 17 languages are given on Map 6.1, together with customised abbreviations and numbers of speakers.



Map 6.1 The 17 languages of the Torres and Banks Islands, in northern Vanuatu

4.3.2 Applying the Comparative Method

Based on primary data I have been collecting on these 17 languages since 1997, I identified regular sound correspondences among them, and created a database of morphological and lexical reconstructions (François 2005, 2013).

The steps involved in applying the Comparative Method should be familiar to historical linguists, since most are also practised with more classical (tree-based) approaches to subgrouping. Data collected in modern languages are analysed in light of regular sound correspondences, so as to identify cognate sets and reconstruct corresponding proto-forms. For each property considered in a given language, it is possible to make reasonably solid hypotheses about whether that property is conservative of earlier stages such as Proto-Oceanic, or results from a local innovation that took place – that is, emerged and diffused – after the initial settlement of Vanuatu.

For example, consider the modern forms for the verb 'steal' in the Torres–Banks languages (ranked geographically from northwest to southeast):

(1) 'steal': HIW *βeneɣ*; LTG *βəneɣ*; LHI *pɒl*; LYP *pɪl*; VLW *ᵐbɛl*; MTP *ᵐbɛl*; LMG *pœl*; VRA *ᵐbɔl*; VRS *ᵐbœl*; MSN *pɔl*; MTA *pal*; NUM *ᵐbal*; DRG *ᵐbaːl*; KRO *ᵐbɛ͡al*; OLR *pal*; LKN *pal*; MRL *ᵐbɐl*.

Knowledge of historical phonology in this region makes it clear that the two Torres forms (HIW *βeneɣ*; LTG *βəneɣ*) are regular reflexes of *panako 'steal', the etymon reconstructed at the level of Proto-Oceanic (Blust 2013). While these two languages exhibit sound change here, they are lexically conservative: these forms thus constitute, for the present purpose, a case of shared retention, and should not count towards subgrouping. By contrast, the forms in the 15 Banks languages all reflect an etymon which can be reconstructed, based on regular sound correspondences, as *ᵐbalu (François 2005: 493). This form is unattested elsewhere in Oceanic, and can therefore safely be flagged as a local lexical innovation shared by the 15 Banks languages. Doing so does not necessarily require positing a unitary 'Proto-Banks' language sharply divided from the rest, like a node in a tree: rather, what is defined here is simply a cluster of 15 communalects which, at some point prior to becoming completely mutually unintelligible, happened to share certain innovations together. (In fact, that cluster is sometimes crosscut by certain isoglosses: see Table 6.1 below.) The identification of innovations requires that variants can be ordered in time. In this case, there is enough evidence to show that *panako predates *ᵐbalu, so the latter is innovative. This procedure sometimes involves reasonings on the relative chronology of changes, whenever this is justified by the data (see François 2011a: 200).

Once each historical innovation was identified following similar procedures, it became possible to indicate which languages reflect it, and which don't. At this point, I deliberately avoided making judgements – which would have been largely arbitrary – regarding whether a given innovation was a 'common' or an 'uncommon' type of change. While this precaution is made necessary by an all-or-nothing approach such as the Tree Model (where an uncommon change can serve as a fatal counter-example to a particular subgrouping hypothesis), it is much less relevant in a model capable of handling innovations in conflicting distributions. In fact, in the event that a subgroup AB were supported by ten 'rare' innovations and BC by ten 'common' ones, there would be no legitimate reason for considering AB to be more strongly supported than BC: the two subgroups should be given equal weight, regardless of the nature (common vs. uncommon) of their internal innovations.

Likewise, I made no attempt to separate shared innovations from changes that potentially could have been innovated independently in two languages (*parallel innovations*), because this too could only be open to speculation. My hypothesis, which proved successful, was that a large enough number of data points should yield a strong genealogical signal based on well supported subgroups, whereas any noise due to parallel innovations would be reduced, due to the low attestation of associated language clusters.

In sum, the key to meaningful results was to first create a large database of historical innovations.

4.3.3 Compiling a database of innovations

I compiled a database of 474 different innovations. These include 21 instances of regular (i.e. systemic) sound change, 116 of irregular (i.e. lexically-specific) sound change, 91 of morphological change, 10 of syntactic change and 236 of lexical replacement.

For each language L and each innovation *i*, the database has '1' when language L reflects *i*; '0' when there is positive evidence that L did not undergo *i*; and a blank whenever the evidence is inconclusive either way. Altogether, the database contains 8058 data points: 2728 positive ('1'), 5040 negative ('0') and 290 agnostic ('–').

Table 6.1 displays a small sample of nine such innovations taken from the database, and shows their distribution across the 17 members of the linkage. Each innovation is here identified using a number (first column) and a label (second column) used here simply as a mnemonic for housekeeping purposes.

Table 6.1 Intersecting isoglosses among Torres and Banks languages: a small sample

Id		HIW	LTG	LHI	LYP	MTP	VLW	LMG	VRA	VRS	MSN	MTA	NUM	DRG	KRO	OLR	LKN	MRL
1	*ᵐbalu	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
2	*late	1	1	1	0	0	0	1	1	1	0	0	0	0	0	0	0	0
3	*suRi	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0
4	*oᵑga	0	0	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
5	*ira	0	0	0	1	1	1	1	1	0	0	0	0	0	0	0	0	0
6	*t>ʔ	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0
7	*one	0	0	0	0	0	0	1	1	1	1	–	1	1	1	0	1	0
8	*wo	0	0	0	0	0	0	0	0	1	1	1	1	1	1	0	0	0
9	*ŋoRo	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1

The nine innovations of Table 6.1 are defined as follows:

1 [*ᵐbalu] LEXICAL REPLACEMENT
POc *panako 'steal' was replaced with a new verb *ᵐbalu (see above)
2 [*late] LEXICALLY-SPECIFIC SOUND CHANGE
*late 'break sth. in two' irregularly changed to *lete
(e.g. VRA *lıʔ* is a regular reflex of *lete but not of *late)
3 [*suRi] MORPHOLOGY
POc verb *suRi 'follow' grammaticalised into a Dative preposition
(e.g. MTP *hij*, a regular reflex of *suRi, encodes Dative: François 2001: 683)
4 [*oᵑga] LEXICALLY-SPECIFIC SOUND CHANGE
POc *waᵑga 'canoe' irregularly changed to *oᵑga
(e.g. MTP *ɔk* is a regular reflex of *oᵑga but not of *waᵑga)
5 [*ira] LEXICALLY-SPECIFIC SOUND CHANGE
*ura 'lobster' (<POc *quraŋ) irregularly changed to *ira (François 2011a: 200)
(e.g. LYP *n-i͡ɛj* is a regular reflex of *ira but not of *ura)
6 [*t > ʔ] REGULAR SOUND CHANGE
*t regularly changed to glottal stop *ʔ*
7 [*one] LEXICALLY-SPECIFIC SOUND CHANGE
*eno 'lie down' (<POc *qenop) metathesised to *one
(e.g. LMG *æn* is a regular reflex of *one but not of *eno)
Note: The etymon *qenop has been lost altogether in Mota, where 'lie down' is a non-cognate form *rsa*. This lexical replacement makes it impossible to empirically assess whether pre-Mota had earlier kept the conservative form *eno (coded as '0') or undergone the metathesis to *one like its neighbours (coded as '1'). Therefore I choose to remain agnostic and mark this language as one where the presence of the innovation cannot be assessed at all (coded as '–'). Historical Glottometry assigns a separate status to such data points, and treats them differently from 0 or 1.
8 [*wo] MORPHOLOGY innovative clitic *wo replaced the NP article *na for alienable non-human nouns (François 2007)
9 [*ŋoRo] LEXICAL REPLACEMENT
POc *matiruR 'sleep' was replaced by *ŋoRo, etymologically 'snore'.

Importantly, all the innovations considered here are unlikely to result from recent borrowing, and can be safely assumed to have been diffused in the earlier times of mutual intelligibility: they are therefore strongly diagnostic of genealogical relations in the sense of the Comparative Method. This is true of cases of lexical replacement selected here, because they involve basic vocabulary items, and because the replacement evidently predated regular sound change in each language (e.g. Lakon has *ŋoRo > *ŋɔː* 'sleep', with regular loss of *R and compensatory lengthening, cf. François 2011b: 150). Instances of *lexically-specific sound change* are also strongly indicative of genealogy, because they are unlikely to diffuse across separate languages: these arbitrary alterations of word forms normally diffuse only across individuals who self-identify as speakers of the same language at the time of the change (Ross 1988: 12; François 2011a: 200).

As the table suggests, plotting innovations on a map of Torres–Banks languages typically yields patterns of intersecting isoglosses, along the lines of Figure 6.3 above. Their linguistic history cannot be represented by a tree, and is better approached using a non-cladistic model.

4.3.4 Displaying results in a NeighborNet

My collaborator Siva Kalyan and I used the database described above as the basis for a number of calculations, in order to identify genealogical subgroups and assess their relative strengths. Figure 6.4 provides a NeighborNet of northern Vanuatu languages, based on rates of pairwise 'acquired similarity' or 'cohesiveness' (number of innovations shared between two languages, as a proportion of the total number of innovations affecting either one).

Because the input data underlying this figure was carefully selected as representing historical *innovations* – not retentions – the distance separating any two languages reflects the accumulation of innovations over time, on one or the other side of a split. In many cases, the language clusters thus defined correspond to genealogical subgroups, as defined by a number of shared innovations. For example, Figure 6.4 reflects the fact that Lemerig belongs simultaneously to two intersecting subgroups: (1) the subgroup *Lehali–Löyöp–Volow–Mwotlap–Lemerig*; and (2) the subgroup *Lemerig–Vera'a–Mwesen–Vurës*.[13] In that sense, NeighborNet offers promising potential for what we are looking for, namely a model for handling and representing intersecting genealogical subgroups.

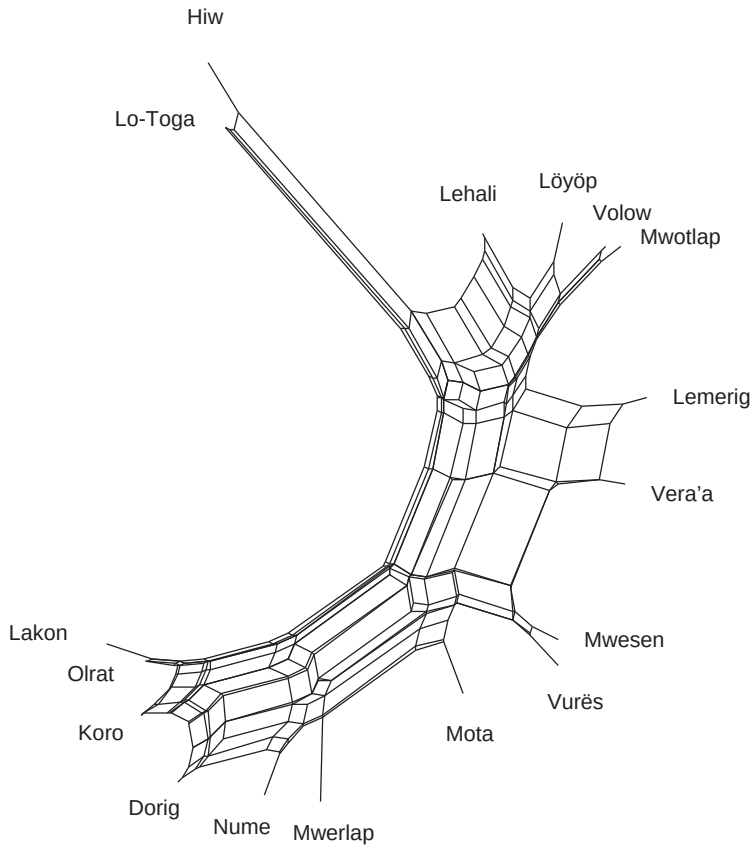


Figure 6.4 A NeighborNet diagram of northern Vanuatu languages, based on rates of acquired similarity

However, NeighborNet has the disadvantage of being ambiguous as to which of the two sides of a split (bundle of parallel lines) corresponds to a genealogical, innovation-defined subgroup. For example, the major split visible between Mota and Mwerlap is indicative of a genealogical subgroup, but doesn't specify which side is innovative:[14] one needs to look up the historical data separately to realise that the relevant subgroup here is the southern one, running from Mwerlap to Lakon.[15] As for the long branch at the top of the figure, it turns out that it encodes one genealogical subgroup on either side: the two Torres languages on the northwestern side (defined by $\varepsilon = 15$ exclusively shared innovations), and the 15 Banks languages to the southeast (with $\varepsilon = 13$); however, this symmetrical structure is not made explicit in the figure.

Furthermore, some of the most prominent splits in this network are actually illusory, because neither side corresponds to any innovation-defined subgroup. For example, the split that runs between Lemerig and Vera'a does not correspond to any isogloss that would encompass either the languages on the northern side (Lemerig to Lo-Toga/Hiw) or those on the southern side (Vera'a to Lakon). In spite of the advantages of NeighborNet, this is, in my view, a major problem if we want to represent genealogical relations in a way that is faithful to the results of the Comparative Method.

4.3.5 The glottometric analysis

The approach developed by Kalyan and François (forthcoming) as Historical Glottometry operates not just on pairs of languages, but on clusters of any size. This is a characteristic it shares with the Tree Model, which also deals with subgroups of various sizes; the only difference is that Historical Glottometry is capable of handling genealogical subgroups even when they intersect.

A genealogical subgroup is defined (section 3.2) as any cluster of languages which have undergone at least one innovation together, at a time when they were still mutually intelligible. In this respect, any historical isogloss potentially defines a subgroup. However, defining subgroups based on weak evidence may run the risk, in some cases, of counting parallel innovations or other accidental resemblances. To avoid this pitfall, Historical Glottometry proposes a method for weighing the amount of evidence supporting each subgroup, so as to reconstruct the most significant patterns in the genealogical history of a language family.

4.3.5.1 Cohesiveness

The absolute number ε of exclusively shared innovations is not the only useful measure of a subgroup's strength. Another way to assess it is to calculate the subgroup's *cohesiveness* (Kalyan and François, forthcoming). This measure (which I have also referred to as 'acquired similarity') is relative rather than absolute: it represents the *proportion* of evidence supporting that subgroup with respect to the entire set of relevant evidence.

For any given subgroup G, let p be the number of supporting innovations (i.e. innovations which include that whole subgroup in their scope, whether exclusively or not), and q the number of conflicting innovations (i.e. innovations whose scope crosscuts G, by involving only some members of G together with some non-members). The total amount of evidence that is relevant for assessing the cohesiveness of G is $(p + q)$. Now, if we call k_G the cohesiveness value of G, we have:

$$k_G = \frac{\textit{number of supporting innovations}}{\textit{total number of relevant innovations}} = \frac{\boldsymbol{p}}{(\boldsymbol{p}+\boldsymbol{q})}$$

Given any cluster of languages, cohesiveness is a measure of how close it is to a perfectly cohesive subgroup. In an ideal tree such as Figure 6.2 above, subgroups are never contradicted by intersecting innovations, and their cohesiveness rate is necessarily 100 per cent. But this rate is hardly ever met with in real-life linkages, where innovations commonly intersect.

The two languages Lemerig and Vera'a share 134 innovations – including $\varepsilon = 9$ which they share exclusively (cf. #6 in Table 1). Conversely, 30 innovations are shared by Lemerig with languages other than Vera'a (cf. #4 in Table 1); and 14 are shared by Vera'a with languages other than Lemerig. In other words, the cohesion of the language pair *Lemerig–Vera'a* is confirmed $p = 134$ times, but betrayed, as it were, $q = 44$ times. The cohesiveness rate of this subgroup is thus $k_{LMG–VRA} = 134/(134 + 44) = 0.75$: this means that, whenever either of its members shared an innovation with at least one other language, then, 75 per cent of the time, the isogloss encompassed both languages, thus confirming this particular subgroup. This figure can be compared with the cohesiveness of the pair *Vera'a–Vurës*, on the same island, which forms a subgroup defined by a single exclusively shared innovation. For this subgroup, $p = 118$ and $q = 88$; so $k_{VRA–VRS} = 118/(118 + 88) = 0.57$: that is, among the many isoglosses that affected either of the languages in this pair, only 57 per cent involved its two members together.

From this comparison, we can make the inference that the ancestors of the Vera'a community had stronger linguistic ties – and by extension, social bonds – with Lemerig to their north ($k_{LMG–VRA} = 75\%$), than with Vurës to their south ($k_{VRA–VRS} = 57\%$) – in spite of the close social ties between today's Vera'a and Vurës communities. Such a metric provides a unique window onto the social networks of the past, based on the traces they left upon modern languages.

4.3.5.2 Subgroupiness

In sum, the degree of support for a genealogical subgroup can be measured in two ways. In absolute terms, its number of *exclusively shared innovations* (ε) indicates the number of times the subgroup is 'attested'; in relative terms, its *cohesiveness* rate (k) indicates how close it is to a perfect subgroup.

These two figures, which constitute equally legitimate measurements of a subgroup's degree of support, are mutually independent. A subgroup for which both k and ε are high is obviously strongly supported: this is the case with the pair *Mwotlap–Volow*, for example, for which $k = 92\%$ and $\varepsilon = 14$. By contrast, the subgroup *Vurës–Mwesen–Mota–Nume–Mwerlap* has both low cohesiveness ($k = 29\%$) and low attestation ($\varepsilon = 2$): it thus has comparatively low support. But certain subgroups are only low on one of these dimensions, and therefore qualify for an intermediate level of support. For example, the pair of languages *Dorig–Koro* has high cohesiveness ($k = 78\%$), but is only attested, in my current database, $\varepsilon = 5$ times. Symmetrically, the whole Banks subgroup – encompassing all 15 languages from Lehali to Lakon – has low cohesiveness ($k = 30\%$), yet is confirmed by many isoglosses ($\varepsilon = 13$).

Ideally, there would be a way to take into account not just one of these two measures, but both of them, as part of an overall assessment of a subgroup's level of support. And indeed, Historical Glottometry proposes to combine ε and k into a single figure: the absolute number of exclusively shared innovations, weighted by the subgroup's cohesiveness. This new metric, called *subgroupiness* (sigma $\varsigma = \varepsilon \times k$), indicates the overall strength of the support for a given subgroup.

Table 6.2 displays subgroupiness values for those northern Vanuatu subgroups which have been mentioned in this chapter.

Table 6.2 Measures of cohesiveness (*k*) and subgroupiness (ς) of a few Torres–Banks subgroups

subgroup	*ε*	*k*	subgroupiness (ς)
MTP-VLW	14	0.92	ς = 14 × 0.92 = **12.82**
HIW-LTG	15	0.83	ς = 15 × 0.83 = **12.45**
LMG-VRA	9	0.75	ς = 9 × 0.75 = **6.75**
DRG-KRO	5	0.78	ς = 5 × 0.78 = **3.90**
whole Banks subgroup	13	0.30	ς = 13 × 0.30 = **3.90**
MRL-NUM-DRG-KRO-OLR-LKN	7	0.43	ς = 7 × 0.43 = **3.00**
LMG-VRA-VRS-MSN	5	0.44	ς = 5 × 0.44 = **2.20**
LHI-LYP-VLW-MTP-LMG	3	0.42	ς = 3 × 0.42 = **1.26**
VRA-VRS	1	0.57	ς = 1 × 0.57 = **0.57**
VRS-MSN-MTA-NUM-MRL	2	0.29	ς = 2 × 0.29 = **0.58**

4.3.5.3 A glottometric diagram

Kalyan and François (forthcoming) calculated subgroupiness rates for all 142 attested subgroups of the Torres–Banks area. Among these, the 32 best supported ones (i.e. those above an arbitrary threshold of $\varsigma \geq 1$), were brought together into a single figure, named a *glottometric diagram* (Figure 6.5). The support for each subgroup is visually represented by having line thickness proportional to *subgroupiness* (ς). The darkness of the contour line is proportional to *cohesiveness* (*k*), with more cohesive subgroups appearing darker.

This result would warrant more commentary than is possible in this chapter; I will stick to the essential aspects. First of all, the subgroupiness values, as well as the diagram derived from them, confirm the statement in section 4.3.1, that northern Vanuatu languages form a *linkage* in which isoglosses, and hence subgroups, constantly intersect. For example, in line with the NeighborNet above, Lemerig [LMG] subgroups both with the four languages to its north (ς = 1.26) *and* with the three languages to its south (ς = 2.20). Similarly Mota [MTA] forms the bridge, as it were, between a northern Banks subgroup (running from Lehali to Mota, ς = 1.03) and a distinct southern Banks subgroup (running from Mota to Lakon, ς = 1.30). The whole island of Gaua, finally, constitutes the epitome of a perfect dialect chain.

It is worthy of notice that the glottometric approach can *also* detect and represent those situations which are 'tree-like' (see section 3.4): for example, Volow and Mwotlap clearly form a subgroup apart from Löyöp; Hiw and Lo-Toga also belong together. Yet evidently, these tree-like patches are a rarity in a language network which is strongly non-tree-like.

While the chaining of languages is essentially coherent with their spatial distribution, a finer grain of observation reveals certain non-trivial patterns that do more than just index geography. For example, even though Volow's location is closer to Mota than to Löyöp (Map 6.1), the position of the three languages in the diagram shows that Volow and Mota are genealogically quite remote (*k* = 36%). Evidently, the ancient society of Mota had very few direct social interactions with its neighbour from Motalava island, and much more with the other islands located to its west – Vanua Lava – or to its south – Gaua, and even the remote Merelava with which Mota forms a genealogical subgroup, in spite of geographic distances. Such results illustrate the potential of the glottometric method for reconstructing the shape of past social networks.

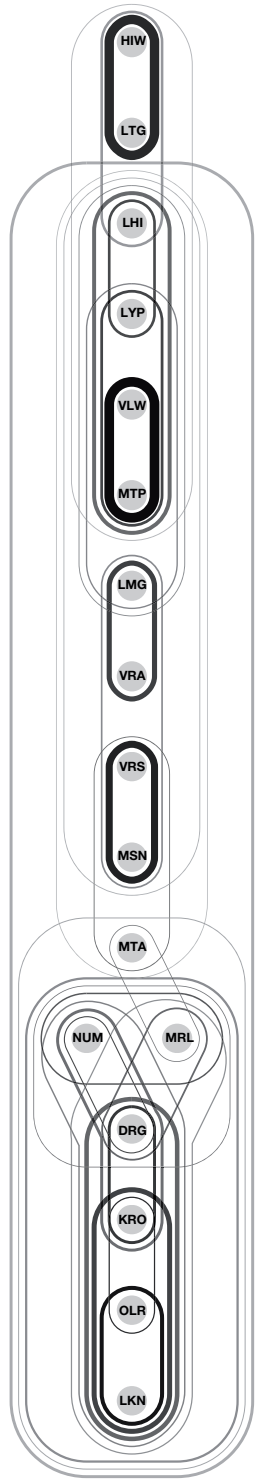


Figure 6.5 A glottometric diagram of the Torres–Banks languages

Glottometric diagrams offer an alternative to the family tree for representing genealogical relations among languages. The analysis of innovations which underlies Historical Glottometry is entirely faithful to the Comparative Method; yet it relies on the Wave Model for one crucial insight, namely that genealogical subgroups may perfectly well crosscut each other. This empirical observation simply reflects the fact that a given community may develop social bonds with several other groups simultaneously.

5 Conclusion

Contrary to widespread belief, there is no reason to think that language diversification typically follows a tree-like pattern, consisting of a nested series of neat splits with loss of contact. Except for the odd case of language isolation or swift migration and dispersal, the normal situation is for language change to involve multiple events of diffusion across mutually intelligible idiolects in a network, typically distributed into conflicting isoglosses. Insofar as these events of language-internal diffusion are later reflected in descendant languages, the sort of language family they define – a 'linkage' – is one in which genealogical relations cannot be represented by a tree, but only by a diagram in which subgroups intersect.

This form of language diversification – probably the most common in the world – requires an approach ultimately inspired by Schmidt's *Wellentheorie* and its overlapping waves. Among various such approaches which have been proposed, Historical Glottometry aims at detecting the genealogical structure of language families in a fine-grained, reliable and testable manner, by combining the strengths of the Comparative Method with a diffusionist, non-cladistic model of language diversification.

Notes

1 I would like to thank Siva Kalyan and Malcolm Ross for their advice on various aspects of the present chapter. This research was presented at the twenty-first International Conference of Historical Linguistics (ICHL21) in Oslo, in August 2013. It forms part of the research strand 'Typology and dynamics of linguistic systems' of the LabEx *Empirical Foundations of Linguistics* (funded by ANR-CGI).

2 See Nettle (1999). For a case study of how these opposing processes interact in a specific region of Melanesia, see François (2012).

3 I follow here the proposal by Haspelmath (2004: 222) to use the term 'genealogical' for what have been traditionally labelled 'genetic' relations, to avoid confusion with biological genetic relations. For a discussion of what is meant by *genealogy* in historical linguistics, see §3.1.

4 There is sometimes ambiguity as to whether social separation is understood as the *cause* or the *consequence* of linguistic divergence. Indeed, social or physical isolation entails that dialects will develop separately from each other; but in addition, following a sort of snowball effect, the more dialects diverge, the higher the language barrier for future communication, and thus the more the social communities may be assumed to develop independently from their neighbours, as their dialects evolve into mutually unintelligible languages. As we will see below, the latter view is quite simplistic, and communication often continues for a long time in spite of earlier events of linguistic divergence.

5 I use the term *diffusion* here in its usual sense of propagation through a social network of individuals (as in Labov 2007). This is distinct from the process of *lexical diffusion*, which describes the way certain forms of sound change propagate across the lexicon (Labov 1994: 421; Krishnamurti 1998).

6 Hale (this volume) expresses a similar idea in terms of individual 'grammars'.

7 While the two processes of diffusion – language-internal vs cross-linguistic – are fundamentally similar in the way they spread through a population, they still differ in their precise psycholinguistic mechanism, and in the nature of the linguistic features they affect (Bowern 2013): for example, 'basic vocabulary' items are more likely to spread through language-internal diffusion than through contact (Haspelmath and Tadmor 2009: 6568). This sort of difference is not relevant to my main point here, which is to say that in both cases, the Tree Model is ill-designed to represent the facts of diffusion – including those that define genealogical relations.

8 For empirical illustrations of this point, see for example Geraghty (1983) for Fijian communalects, Garrett (2006) for ancient Greek dialects, François (2011a: 201) for northern Vanuatu.

9 Societies differ on how much linguistic fragmentation they tolerate. Some more centralised societies may involve a higher degree of levelling between dialects, in such a way that a change affecting the more central or influential varieties will rapidly spread to the whole network of individuals who self-identify as speakers of that 'language'. Conversely, some societies are more tolerant towards internal diversity, and exert less pressure towards dialect levelling.

10 In some cases, dialect levelling may erase the earlier entangled structure of a continuum, and produce the 'mirage' of discrete subgroups (Garrett 2006). For example, in Figure 6.3 above, should dialects E and F be wiped out as distinct varieties, then the isoglosses would appear nested again, and the family could be rendered by a tree. However, a tree-like structure is not a necessary result of dialect levelling. Thus, if the process meant the demise of dialects B, C and H in Figure 6.3 but the survival of other varieties, then the genealogical structure of the linkage descended from this continuum would still resist any cladistic approach. For example, it can be shown that Italian, Spanish and French do not properly fit into a tree, even without considering the numerous intervening dialects (Kalyan and François forthcoming).

11 Another problem is that some of the work conducted using these methods is not based on the Comparative Method. Dunn *et al.* (2008), for example, identify their subgroups based on a matrix of typological features such as word order, rather than on linguistic reconstruction and the identification of innovations.

12 Because dialectologists use the term 'isogloss' regardless of its historical nature, one may want to specify that the isoglosses used in Historical Glottometry are all HISTORICAL ISOGLOSSES – *à la* Bloomfield (1933: 316) or Anttila (1989: 305).

13 Among other relevant diagnostic innovations, the first of these two subgroups is defined by the lexically-specific change *wa^{ŋ}ga > *o^{ŋ}ga 'canoe' (see Table 6.1); the second by the lexically-specific dissimilation *mama^{n}ri^{n}ri > *mamaɣi^{n}ri 'cold'.

14 This limitation could be rectified by including the proto-language, Proto-Oceanic, as one of the taxa displayed in the NeighborNet. In this case, whichever side of the split does *not* include the proto-language, would be the group defined by innovations. The practice of including the ancestral node as a taxon in a NeighborNet, however, does not seem to be widely followed in linguistics.

15 For example, these six languages share the use of a preposition *ma^{ŋ}ge 'above'; or the lexically-specific loss of the phoneme *R in *nañoRap 'yesterday' and *waRisa 'two days from now' (François 2011b: 157).

Further reading

Croft, William. 2000. *Explaining language change: an evolutionary approach*. London: Longman.

Goebl, Hans. 2006. Recent advances in Salzburg dialectometry. *Literary and Linguistic Computing* 21(4): 411–435.

Heggarty, Paul, Warren Maguire and April McMahon. 2010. Splits or waves? Trees or webs? How divergence measures and network analysis can unravel language histories. *Philosophical Transactions of the Royal Society B: Biological Sciences* 365: 3829–3843.

Milroy, James and Lesley Milroy. 1985. Linguistic change, social network and speaker innovation. *Journal of Linguistics* 21(2): 339–384.

Ross, Malcolm. 1997. Social networks and kinds of speech-community event. In Roger Blench and Matthew Spriggs, *Archaeology and language*. Volume 1: *Theoretical and methodological orientations*. London: Routledge, 209–261.

References

Aikhenvald, Alexandra and R. M. W. Dixon (eds) 2001. *Areal diffusion and genetic inheritance: problems in comparative linguistics*. Oxford: Oxford University Press.

Anttila, Raimo. 1989. *Historical and Comparative Linguistics*. Amsterdam/Philadelphia: John Benjamins.

Biggs, Bruce. 1965. Direct and indirect inheritance in Rotuman. *Lingua* 14: 383–415.

Bloomfield, Leonard. 1933. *Language*. New York: Holt.

Blust, Robert. 2013. *Austronesian Comparative Dictionary*, web edition [Available at: http://www.trussel2.com/ACD; accessed 22 May 2013].

Bossong, Georg. 2009. Divergence, convergence, contact. Challenges for the genealogical classification of languages. In Kurt Braunmüller and Juliane House, *Convergence and divergence in language contact situations*. Amsterdam/Philadelphia: John Benjamins, 13–40.

Bowern, Claire. 2006. Another look at Australia as a linguistic area. In Yaron Matras, April McMahon and Nigel Vincent (eds) *Linguistic areas. Convergence in historical and typological perspective*. Basingstoke: Palgrave Macmillan, 244–265.

——2013. Relatedness as a factor in language contact. *Journal of Language Contact* 6: 411–432.

Bowern, Claire and Quentin Atkinson. 2012. Computational phylogenetics and the internal structure of Pama-Nyungan. *Language* 88(4): 817–845.

Bryant, David, Flavia Filimon and Russell D. Gray. 2005. Untangling our past: languages, trees, splits and networks. In Ruth Mace, Clare J. Holden and Stephen Shennan (eds) *The evolution of cultural diversity: phylogenetic approaches*. London: UCL Press, 67–83.

Campbell, Lyle. 2004. *Historical linguistics: an introduction*. Edinburgh: Edinburgh University Press.

Campbell, Lyle and William J. Poser. 2008. *Language classification: history and method*. Cambridge: Cambridge University Press.

Chambers, Jack K. and Peter Trudgill. 1998. *Dialectology*. Cambridge: Cambridge University Press.

Chappell, Hilary. 2001. Language contact and areal diffusion in Sinitic languages: problems for typology and genetic affiliation. In Aikhenvald and Dixon (eds), 328–357.

Croft, William. 2000. *Explaining language change: an evolutionary approach*. London: Longman.

Dixon, R. M. W. 1997. *The rise and fall of languages*. Cambridge: Cambridge University Press.

Drinka, Bridget. 2013. Phylogenetic and areal models of Indo-European relatedness: The role of contact in reconstruction. *Journal of Language Contact* 6(2): 379–410.

Dunn, Michael, Stephen Levinson, Eva Lindström, Ger Reesing and Angela Terrill. 2008. Structural phylogeny in historical linguistics: methodological explorations applied in Island Melanesia. *Language* 84(4): 710–759.

Durie, Mark and Malcolm Ross (eds). 1996. *The Comparative Method reviewed: regularity and irregularity in language change*. Oxford: Oxford University Press.

Enfield, Nicholas. 2003. *Linguistic epidemiology: semantics and grammar of language contact in mainland Southeast Asia*. London: Routledge-Curzon.

——2008. Transmission biases in linguistic epidemiology. *Journal of Language Contact* Thema 2: 299–310.

Ernst, Gerhard, Martin-Dietrich Gleßgen, Christian Schmitt and Wolfgang Schweickard (eds) 2009. *Romanische Sprachgeschichte – Ein Internationales Handbuch zur Geschichte der Romanischen Sprachen/Histoire linguistique de la Romania – Manuel international d'histoire linguistique de la Romania*. Berlin: De Gruyter Mouton.

Forster, Peter, Alfred Toth and Hans-Jürgen Bandelt. 1998. Evolutionary network analysis of word lists: visualising the relationships between Alpine Romance languages. *Journal of Quantitative Linguistics* 5(3): 174–187.

Fox, Anthony. 1995. *Linguistic reconstruction: an introduction to theory and method*. Oxford: Oxford University Press.

François, Alexandre. 2001. Contraintes de structures et liberté dans l'organisation du discours: une description du mwotlap, langue océanienne du Vanuatu. PhD dissertation. Université Paris-IV Sorbonne, Paris.

——2005. Unraveling the history of the vowels of seventeen northern Vanuatu languages. *Oceanic Linguistics* 44(2): 443–504.

——2007. Noun articles in Torres and Banks languages: conservation and innovation. In Jeff Siegel, John Lynch and Diana Eades (eds) *Language description, history and development: linguistic indulgence in memory of Terry Crowley*. New York: Benjamins, 313–326.

——2011a. Social ecology and language history in the northern Vanuatu linkage: a tale of divergence and convergence. *Journal of Historical Linguistics* 1(2): 175–246.

——2011b. Where *R they all? The geography and history of *R loss in Southern Oceanic languages. *Oceanic Linguistics* 50(1): 140–197.

——2012. The dynamics of linguistic diversity. Egalitarian multilingualism and power imbalance among northern Vanuatu languages. *International Journal of the Sociology of Language* 214: 85–110.

——2013. Shadows of bygone lives: the histories of spiritual words in northern Vanuatu. In Robert Mailhammer (ed.) *Lexical and structural etymology: beyond word histories*. Berlin: DeGruyter Mouton, 185–244.

Garrett, Andrew. 2006. Convergence in the formation of Indo-European subgroups: phylogeny and chronology. In Peter Forster and Colin Renfrew (eds) *Phylogenetic methods and the prehistory of languages*. Cambridge: McDonald Institute for Archaeological Research, 139–151.

Geraghty, Paul A. 1983. *The history of the Fijian languages*. Honolulu: University of Hawaii Press.

Giles, Howard and Tania Ogay. 2007. Communication accommodation theory. In Bryan B. Whaley and Wendy Samter (eds) *Explaining communication: contemporary theories and exemplars*. London: Routledge, 293–310.

Gilliéron, Jules. 1880. *Petit Atlas phonétique du Valais roman (sud du Rhône)*. Paris: Champion.

Goebl, Hans. 2006. Recent advances in Salzburg dialectometry. *Literary and Linguistic Computing* 21(4): 411–435.

Gray, Russell D., David Bryant and Simon J. Greenhill. 2010. On the shape and fabric of human history. *Philosophical Transactions of the Royal Society London, B* 365: 3923–3933.

Gray, Russell D., Alexei J. Drummond and Simon J. Greenhill. 2009. Language phylogenies reveal expansion pulses and pauses in Pacific settlement. *Science* 323 (5913): 479–483.

Greenhill, Simon J., Alexei J. Drummond and Russell D. Gray. 2010. How accurate and robust are the phylogenetic estimates of Austronesian language relationships? *PLoS One* 5 (3): e9573.

Greenhill, Simon J. and Russell D. Gray. 2009. Austronesian language phylogenies: myths and misconceptions about Bayesian computational methods. In Alexander Adelaar and Andrew Pawley (eds) *Austronesian historical linguistics and culture history: a festschrift for Robert Blust*. Canberra: Pacific Linguistics, 375–397.

Guarisma, Gladys and Wilhelm J.G. Möhlig (eds) 1986. *La méthode dialectométrique appliquée aux langues africaines*. Berlin: Reimer.

Hashimoto, Mantaro. 1992. Hakka in Wellentheorie perspective. *Journal of Chinese Linguistics* 20: 1–49.

Haspelmath, Martin. 2004. How hopeless is genealogical linguistics, and how advanced is areal linguistics? *Studies in Language*, 28(1): 209–223.

Haspelmath, Martin and Uri Tadmor. 2009. *Loanwords in the world's languages: a comparative handbook*. Berlin: Mouton de Gruyter.

Heggarty, Paul, Warren Maguire and April McMahon. 2010. Splits or waves? Trees or webs? How divergence measures and network analysis can unravel language histories. *Philosophical Transactions of the Royal Society B: Biological Sciences* 365: 3829–3843.

Hock, Hans Henrich. 1991. *Principles of historical linguistics*. Berlin: de Gruyter.

Holton, Gary. 2011. A geo-linguistic approach to understanding relationships within the athabaskan family. Paper read at the international workshop *Language in Space: Geographic Perspectives on Language Diversity and Diachrony*. Boulder, Colorado.

Huehnergard, John and Aaron Rubin. 2011. Phyla and Waves: models of classification. Semitic languages. In Stefan Weninger, Geoffrey Khan, Michael P. Streck and Janet E. Watson (eds) *Semitic languages: an international handbook*. Berlin: de Gruyter Mouton, 259–278.

Kalyan, Siva, and Alexandre François. forthcoming. Freeing the Comparative Method from the Tree Model: a framework for Historical Glottometry. In Ritsuko Kikusawa and Lawrence Reid (eds) *Let's talk about trees: tackling problems in representing phylogenic relationships among languages*. Osaka: National Museum of Ethnology.

Krauss, Michael E. and Victor Golla. 1981. Northern Athabaskan languages. In June Helm (ed.) *Handbook of North American Indians*. Volume 6: *Subarctic*. Washington DC: Smithsonian Institute, 67–85.

Krishnamurti, Bh. 1998. Regularity of sound change through lexical diffusion: A study of s > h > Ø in Gondi dialects. *Language Variation and Change* 10: 193–220.

Labov, William. 1963. The social motivation of sound change. *Word* 19: 273–309.

——1994. *Principles of linguistic change.* Volume 1: *Internal factors*. Oxford: Blackwell.

——2001. *Principles of linguistic change.* Volume 2: *Social factors*. Oxford: Blackwell.

——2007. Transmission and diffusion. *Language* 83(2): 344–387.

Leskien, August. 1876. *Die Declination im Slawisch-Litauischen und Germanischen*. Leipzig: Hirzel.

Lynch, John. 2000. Linguistic subgrouping in Vanuatu and New Caledonia. In Bill Palmer and Paul A. Geraghty (eds) *Proceedings of the Second International Conference on Oceanic Linguistics (SICOL).* Volume 2: *Historical and descriptive studies*. Canberra: Pacific Linguistics, 155–184.

Matras, Yaron and Peter Bakker (eds) 2003. *The mixed language debate: theoretical and empirical advances*. Berlin: Walter de Gruyter.

Milroy, James and Lesley Milroy. 1985. Linguistic change, social network and speaker innovation. *Journal of Linguistics* 21(2): 339–384.

Milroy, Lesley. 1987. *Language and social networks*. Oxford: Blackwell.

Nerbonne, John. 2010. Measuring the diffusion of linguistic change. *Philosophical Transactions of the Royal Society B: Biological Sciences* 365: 3821–3828.

Nettle, Daniel. 1999. *Linguistic diversity*. Oxford: Oxford University Press.

Page, Roderick D. M. and Edward C. Holmes. 2009. *Molecular evolution: a phylogenetic approach.* Oxford: Blackwell.

Pawley, Andrew. 1999. Chasing rainbows: implications of the rapid dispersal of Austronesian languages for subgrouping and reconstruction. In Elizabeth Zeitoun and Paul J.-K. Li (eds) *Selected Papers from the Eighth International Conference on Austronesian Linguistics*. Taipei: Academia Sinica, 95–138.

Pawley, Andrew K. and Malcolm D. Ross. 1995. The prehistory of Oceanic languages: a current view. In Peter S. Bellwood, James J. Fox and Darrell T. Tryon (eds) *The Austronesians: historical and comparative perspectives*. Canberra: Australian National University, 39–80.

Penny, Ralph John. 2000. *Variation and change in Spanish*. Cambridge: Cambridge University Press.

Pulgram, Ernst. 1961. The nature and use of proto-languages. *Lingua* 10: 18–37.

Ramat, Paolo. 1998. The Germanic languages. In Anna Giacalone Ramat and Paolo Ramat (eds) *The Indo-European languages*. London: Routledge, 380–414.

Rankin, Robert. 2003. The Comparative Method. In Brian D. Joseph and Richard D. Janda (eds) *The handbook of historical linguistics*. Oxford: Blackwell, 183–212.

Romaine, Suzanne. 1988. *Pidgin and creole languages*. London: Longman.

Ross, Malcolm. 1988. *Proto Oceanic and the Austronesian languages of western Melanesia*. Canberra: Pacific Linguistics.

——1996. Contact-induced change and the Comparative Method: cases from Papua New Guinea. In Durie and Ross (eds), 180–217.

——1997. Social networks and kinds of speech-community event. In Roger Blench and Matthew Spriggs, *Archaeology and language.* Volume 1: *Theoretical and methodological orientations*. London: Routledge, 209–261.

——2001. Contact-induced change in Oceanic languages in North-West Melanesia. In Aikhenvald and Dixon (eds), 134–166.

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

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

Schuchardt, Hugo. 1885. *Über die Lautgesetze: Gegen die Junggrammatiker*. Berlin: Oppenheim.

Séguy, Jean. 1973. La dialectométrie dans l'Atlas linguistique de la Gascogne. *Revue de Linguistique Romane* 145–146: 1–24.

Siegel, Jeff. 2008. *The emergence of pidgin and creole languages*. Oxford: Oxford University Press.

Southworth, Franklin C. 1964. Family-tree diagrams. *Language* 40(4): 557–565.

Street, Richard L. and Howard Giles. 1982. Speech accommodation theory: a social cognitive approach to language and speech behavior. In Michael E. Roloff and Charles R. Berger (eds) *Social cognition and communication*. Beverly Hills: Sage, 193–226.

Szmrecsanyi, Benedikt. 2011. Corpus-based dialectometry: a methodological sketch. *Corpora* 6(1): 45–76.

Toulmin, Matthew. 2006. Reconstructing linguistic history in a dialect continuum: the Kamta, Rajbanshi, and Northern Deshi Bangla subgroup of Indo-Aryan. PhD dissertation. The Australian National University, Canberra.

——2009. *From linguistic to sociolinguistic reconstruction: the Kamta historical subgroup of Indo-Aryan*. Canberra: Pacific Linguistics.

Trudgill, Peter. 1986. *Dialects in contact*. Oxford: Blackwell.

Tryon, Darrell. 1996. Dialect chaining and the use of geographical space. In Joël Bonnemaison, Kirk Huffman, Christian Kaufmann and Darrell Tryon (eds) *Arts of Vanuatu*. Bathurst: Crawford House Press, 170–181.

van Driem, George. 2001. *Languages of the Himalayas: an ethnolinguistic handbook of the greater Himalayan region*. Leiden: Brill.

Wenker, Georg. 1881. *Sprachatlas von Nord- und Mitteldeutschland: Text und Einleitung. Auf Grund von systematisch mit Hülfe der Volksschullehrer gesammeltem Material aus circa 30000 Orten*. Strasbourg, London: Trübner.