Freeing the Comparative Method from the tree model

A framework for Historical Glottometry

to appear in:
Let's talk about trees: Tackling Problems in Representing Phylogenic Relationships among Languages, ed. by R. Kikusawa & L. Reid (Senri Ethnological Studies). Osaka: National Museum of Ethnology.

Siva Kalyan
Northumbria University
— Australian National University

Alexandre François
CNRS LaCiTO; Labex Empirical Foundations of Linguistics
— Australian National University

Abstract

Since the beginnings of historical linguistics, the family tree has been the most widely accepted model for representing historical relations between languages. While this sort of representation is easy to grasp, and allows for a simple, attractive account of the development of a language family, the assumptions made by the tree model are applicable in only a small number of cases: namely, when a speaker population undergoes successive splits, with subsequent loss of contact among subgroups. A tree structure is unsuited for dealing with dialect continua, as well as language families that develop out of dialect continua (for which Ross 1988 uses the term "linkage"); in these situations, the scopes of innovations (in other words, their isoglosses) are not nested, but rather they persistently intersect, so that any proposed tree representation is met with abundant counterexamples. In this paper, we define "Historical Glottometry", a new method capable of identifying and representing genealogical subgroups even when they intersect. Finally, we apply this glottometric method to a specific linkage, consisting of 17 Oceanic languages spoken in northern Vanuatu.

1. Introduction

The use of genealogical trees for the representation of language families is nearly as old as the discipline of historical linguistics itself; it was first proposed by August Schleicher in 1853, six years before Darwin proposed a tree model in evolutionary biology (e.g. Minaka & Sugiyama 2012: 177). It has since been the dominant method of visualising historical relationships among languages, and for good reason: its simple structure allows any hypothetical representation of a language family to be interpreted unambiguously as a set of claims about the sequence of demographic and social events that actually occurred in the history of the communities involved. These hypotheses can then potentially be falsified by new data or analysis, leading to a more valid representation. Other methods of representing the historical relationships among languages have from time to time been proposed and defended—e.g. Johannes Schmidt's (1872) "Wave Model", Southworth's (1964) "tree-envelopes" (akin to the "population trees" used in phylogeography, e.g. Avise 2000: 32), Anttila's (1989: 305) isogloss map, Hock's (1991: 452) "'truncated octopus'-like tree", van Driem's (2001) "fallen leaves", and most recently NeighborNet (Bryant *et al.* 2005), among many others. However, to our knowledge, none of these has combined precision and formalisation with direct interpretability in terms of historical events, to the extent that has been achieved by the family-tree model.[1]

Yet there are important reasons to be dissatisfied with the family-tree model (as has frequently been pointed out; see also Bloomfield 1933: §§18.9–12). In particular, it rests entirely on the assumption that the process of language diversification is one where language communities undergo successive splits—via migration or other forms of social disruption—with subsequent loss of contact. While this particular social scenario may have occurred occasionally (e.g. in the separation of Proto-Oceanic from the remainder of the Austronesian language family; see Pawley 1999), it can hardly be regarded as the general case.

The way language change arises is via a process of *language-internal diffusion* (François forthc.; cf. Labov 1963, Milroy & Milroy 1985, Croft 2000: 166–195; Enfield 2008)—as speakers in a network imitate each other so as to jointly adopt an innovative speech habit. When the innovation settles into a certain section of the social group, it becomes part of its linguistic heritage and can be transmitted to its descendants. This diffusion process is the underlying mechanism behind "genetic" relations (or better, to use Haspelmath's (2004:222) preferred term, "genealogical" relations) among languages, whereby each subgroup is defined by the innovations its members have undergone together. Whereas contact-induced change takes place between separate languages, the process of *language-internal diffusion* that defines language genealogy involves mutually intelligible speech varieties.

The tree model can represent genealogical relations in just one particular case: when a language community has split into separate groups, each of which later goes through its own innovations. But this model cannot properly handle the frequent case when adjacent speech communities remain in contact even after undergoing innovations that increase their difference. In such situations, provided the speech varieties remain mutually intelligible for some time, nothing prevents successive innovations from targeting overlapping portions of the network: e.g. one isogloss targeting dialects A-B-C, another one C-D-E, then B-C, then D-E-F, etc. In such cases of dialect chains or networks, frequently observed in dialectology (and described further below), the layering of partially overlapping innovations results in *intersecting* genealogical subgroups—a situation which cannot be addressed by the tree model (Gray *et al.* 2010:3229).

[1] The authors wish to thank Malcolm Ross, Mark Donohue and Martine Mazaudon for their comments on an earlier draft of this paper. They would also like to thank the other participants of the

As is increasingly evident from the work of a number of historical linguists, this sort of intersecting configuration typical of dialect continua is also the normal situation in most language families around the world (e.g. Geraghty 1983; Ross 1988; Toulmin 2009; Heggarty *et al.* 2010; Huehnegaard & Rubin 2011). Of course, any set of data could be forced into a tree structure, but in most cases this can only be done by selectively discarding some of the data—no doubt in good faith—so as to retain only those which are compatible with a particular subgrouping hypothesis. Debates about which tree best represents the language family thus usually boil down to (often pointless) arguments over which parts of the data may be ignored.

In this study, we start by elaborating on the arguments and the claims made in the preceding paragraphs, by illustrating in greater detail how trees are used in historical linguistics, and discussing their advantages and disadvantages. We then move to the task of proposing a new method of representing genealogical relationships among languages, which we call Historical Glottometry. While ultimately inspired by the Wave Model which Schmidt (1872) proposed as an alternative to the family tree, our method also draws on the quantitative approach of dialectometry (Séguy 1973; Goebl 2006; Szmrecsányi 2011). We hope this model provides more realistic insights into language history than the tree model, while still combining precision and formalisation with historical interpretability. Finally, we illustrate our model by applying it to a group of seventeen Oceanic languages spoken in Vanuatu, an archipelago in the south Pacific.

2. Subgrouping in the tree model

2.1. An example from Indo-European

Consider the family tree shown in Figure 1, which represents a selection from the family of Indo-European languages. At the bottom are languages that are currently spoken; languages higher in the tree are ancestors of the languages that branch from them. Each nodal ancestor is called a proto-language, whose descendants together form a subgroup.

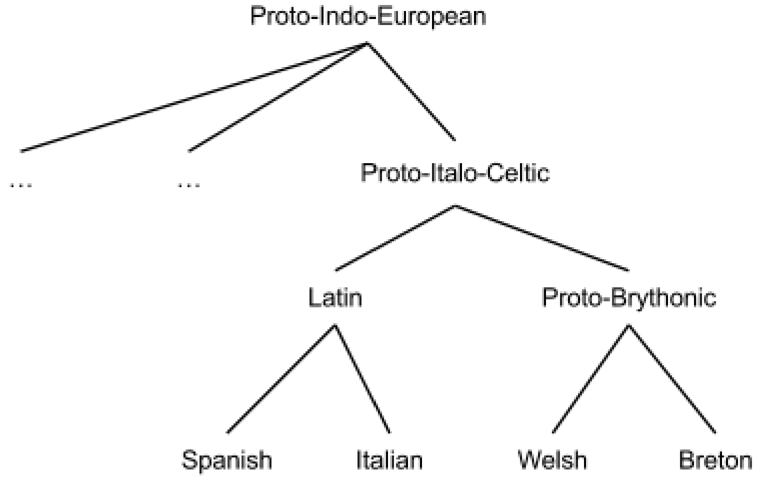


Figure 1: A selection of Indo-European languages, organised as a tree.

In some cases, ancestor languages have been preserved in writing; thus we have direct evidence that (some variety of) Latin is the common ancestor of Spanish and Italian. In other cases, the ancestors are hypothetical, and must be reconstructed by comparing their surviving descendants; thus it is merely a hypothesis that there was a unified Proto-Brythonic language from which Welsh and Breton are descended, and the features of this proto-language are also hypothetical.

Ancestral languages (whether attested or reconstructed) can themselves be compared, and their own ancestors hypothesised and reconstructed, in a recursive fashion. Thus, some linguists (e.g. Kortlandt 2007) believe that Latin and Proto-Brythonic ultimately descend from a language termed Proto-Italo-Celtic (PIC).[2] Repeatedly applying this process of comparison and reconstruction—called the *Comparative Method*—leads to proto-languages further and further back in time, ultimately ending in Proto-Indo-European (PIE).[3]

Granted that the uppermost node, Proto-Indo-European, is valid (since the Indo-European languages are indeed related to one another), on what basis are lower-level proto-languages (or equivalently, subgroups) posited? For example, why isn't Welsh grouped with Latin, separately from Breton, as in the fictitious Figure 2?

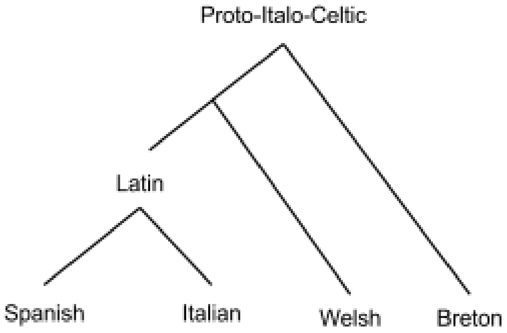


Figure 2: An incorrect tree of Italo-Celtic languages.

The reason is that this would imply that Latin and Welsh both exhibit certain changes (or *innovations*) from PIC (and hence, from PIE) that are not exhibited by Breton. But there are no notable innovations of this kind. Also, Figure 2 would imply that there are *no* innovations shared by Welsh and Breton which are not also shared by Latin (and all other members of the Italo-Celtic subgroup). This too is false: for example, the Brythonic languages changed **k^w* to *p*, and changed **s* to *h* at the beginnings of words (Schmidt 2002: 80–81); Latin, on the other hand, preserved these sounds intact. In sum, the representation in Figure 1 is more faithful to the empirical data we have from attested languages, than is Figure 2.

As we have just illustrated, in the Comparative Method, a subgroup is posited on the basis of *exclusively shared innovations* among its members—a principle first formulat-

[2] Brythonic is actually a branch of Celtic, which in turn is a branch of Italo-Celtic; likewise, Latin is a member of the Italic branch of Italo-Celtic. The fact that the existence of Proto Italo-Celtic is controversial is irrelevant to the present demonstration—what is important is that Latin and the Brythonic languages do in fact have a common ancestor (even if that ancestor turns out to be nothing other than Proto-Indo-European itself).

[3] On general principles of the comparative method, see Hock (1991), Campbell (2004), Crowley & Bowern (2010), among many others.

ed by Leskien (1876: xiii). In other words, a subgroup represents a hypothesis that all of its members share certain innovations that are not exhibited by any other language, and that any innovation that a member shares with a non-member is necessarily shared by *all* members. (This is similar to how, in phylogenetics, clades are interpreted as monophyletic groups defined by synapomorphies: see Skelton et al. 2002: 27–28.)

Let us now consider what happens when we add another language—French—to our tree. There is no question but that French is a descendant of Latin; hence it should ultimately be a daughter of the "Latin" node. However, there are multiple ways in which it could be put into a tree together with Spanish and Italian (Figure 3). Which of these choices is correct?

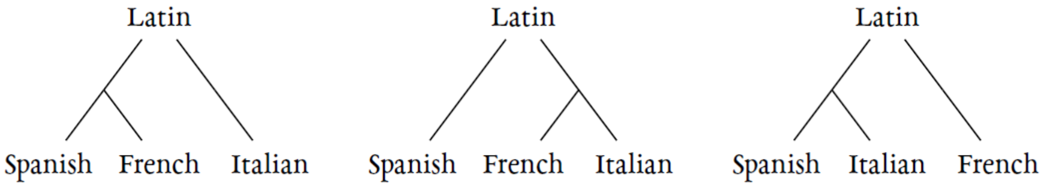


Figure 3: Three possible ways to represent the relations between Spanish, French and Italian.

Choice 1, with Spanish and French forming a subgroup, seems justified by the innovations that are shared between these two languages, and not shared by Italian: for example, the irregular change of *a(u)scultāre* 'listen' to **escultāre* > Sp. *escuchar*, Fr. *écouter*, vs. It. *ascoltare* (Berger & Brasseur 2004: 90); intervocalic lenition of **p*—e.g. *rīpa* 'riverbank' > Sp. *riba*, Fr. *rive*, vs. It. *ripa* (Posner 1996: 234); and the palatalisation of **ct* clusters—e.g. *factum* 'done' > Sp. *hecho*, Fr. *fait* vs. It. *fatto* (Hall 1950: 25). However, one can also find innovations shared by French and Italian but not by Spanish, which would argue in favour of choice 2: for example, the innovative weak past participle suffix **-ūtus* which affected many verbs—e.g. **sapūtus* 'known' > It. *saputo*, Fr. *su*, as opposed to Sp. *sabido* < **sapītus* (Alkire & Rosen 2010: 177); or numerous lexical innovations such as **diurnu* > It. *giorno*, Fr. *jour* 'day', replacing Lat. *diēs* (Sp. *día*), or **manducāre* 'chew' > It. *mangiare*, Fr. *manger* 'eat', replacing Lat. *comedere* (Sp. *comer*). Finally, one could cite evidence in favour of subgrouping Spanish and Italian together as opposed to French (as in choice 3), e.g. the irregular change of Lat. *ecce* to **accu* (Wüest 1994), as in the (feminine) distal demonstrative **accu-illa* > Sp. *aquella*, It. *quella*, where French preserves *ecce* (**ecce-illa* > Fr. *celle*); or the irregular insertion of /n/ in *hibernum* 'winter', yielding **inbernu* > Sp. *invierno*, It. *inverno*, vs. Fr. *hiver* (Alkire & Rosen 2010: 339). Many other examples of exclusively shared innovations[4] could be found for each of the three language pairs. In all cases, the nature of the changes (especially phonological and morphological change, whether regular or irregular) is typical of the sort of evidence that is traditionally considered diagnostic of genealogical subgroups under the Comparative Method.

[4] Obviously, the term "exclusively" must be understood within the restricted set of three languages taken here for the sake of discussion. Some of the innovations shared by French and Spanish are also shared with Catalan, Portuguese, etc., but this is not relevant for the present demonstration. (Interestingly, Catalan seems to exhibit most of the innovations mentioned.)

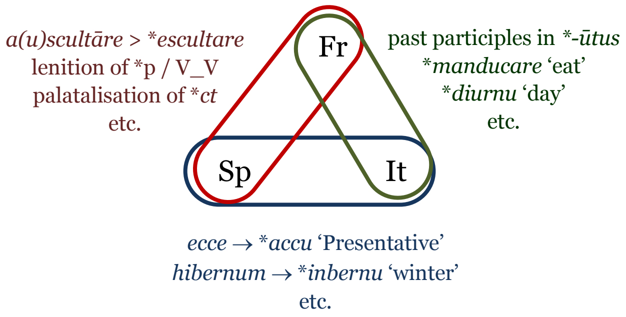

a(u)scultāre > *escultare
lenition of *p / V_V
palatalisation of *ct
etc.
Fr
past participles in *-ūtus
*manducare 'eat'
*diurnu 'day'
etc.
Sp
It
ecce → *accu 'Presentative'
hibernum → *inbernu 'winter'
etc.

Figure 4: Historical evidence supports three intersecting subgroups involving Spanish, French and Italian—a situation incompatible with the family tree model.

In this particular case, the data simultaneously support three intersecting subgroups (Figure 4): Spanish–French, French–Italian and Spanish–Italian. The tree model, which would force us to privilege one of these three groupings at the expense of the other two,[5] is unable to do justice to the empirical evidence.

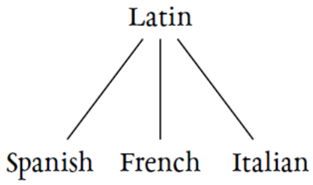


Figure 5: A rake (or "polytomy").

One could be tempted to represent this thorny situation by resorting to the diagram in Figure 5, which does not necessarily commit us to any subgrouping hypothesis. This sort of diagram (cf. Ross 1997: 213) is sometimes used as an "agnostic" representation, which Pawley (1999) calls a "rake-like" structure, and van Driem (2001) likens to "fallen leaves". (In phylogenetics this is known as "(soft) polytomy":[6] see Page & Holmes 2009: 13.) Yet it too is unsatisfactory, as it could be interpreted as claiming that there are *no* exclusively shared innovations between Spanish and French, between French and Italian, or between Spanish and Italian, when—as we have seen—there is in fact solid, positive evidence for all of these. (In phylogenetic terms, a rake is ambiguous between "soft polytomy" and "hard polytomy".) Even if we specifically exclude this latter interpretation, we are only left with the impression that science is simply incapable of unraveling the precise linguistic history of the language family. While this is sometimes the case due to lack of data, it is certainly not the case in such a well-documented family as Romance.

The history of individual changes across Romance dialects and languages is extremely well-known: if this family cannot be represented by a tree, then this cannot be

[5] This is what Hall (1950) does: his assumption that languages must evolve following a cladistic model has him force the data into a tree structure. His "Western Romance" node, by grouping French and Spanish together, arbitrarily favours only one of the three groupings outlined here, and deliberately ignores any conflicting evidence.

[6] We are grateful to Nobuhiro Minaka (p.c.) for pointing this out.

due to a lack of data, but to the inherent flaws of the tree model itself: in particular, the axiom that genealogical subgroups defined by exclusively shared innovations are necessarily nested, and never intersect. This axiom results from an incorrect understanding of language change (cf. Bossong 2009, François forthc.), namely that an innovation consistently results in total social isolation and lack of contact with communities that did not undergo the innovation—an incorrect assumption in most of the world's history. What we see, on the contrary, is that the spread of an innovation within part of a dialect network, insofar as it still allows mutual intelligibility with non-participating dialects, can perfectly well be followed by other innovations whose geographical scope may cross-cut its own, resulting in intersecting subgroups. We need a model of language relationships that is capable of accommodating such situations in a more accurate and faithful way than the tree model.

2.2. The problem of linkages

We can generalise our observations above by considering an abstract case, consisting of a family of three languages: A, B, and C. If A and B have some exclusively shared innovations, but neither B and C nor A and C do, then the situation is amenable to a tree representation (as in choice 1 in Figure 3 above). Historically, this represents a situation where the Proto-ABC speech community somehow split into two groups, one of which (the common ancestor of the modern A and B communities) underwent certain linguistic innovations, separately from C; these innovations are said to have resulted in a hypothetical language "Proto-AB". Later on, a similar split took place in the Proto-AB community, that resulted in the separate development of A and B.

But another situation is also possible, as we saw in the case of Romance languages. This is the case where there are exclusively shared innovations not only between A and B, but also between B and C, and/or between A and C: that is, a situation in which shared innovations define intersecting groupings—see Figure 6 (and Figure 4 above).

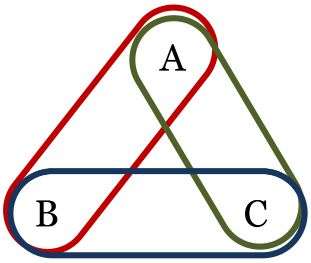


Figure 6: When shared innovations intersect

This situation cannot be represented using the tree model, which assumes that a language can belong to one genealogical subgroup only. The only way to force the data into a tree—and posit, for example, a subgroup AB—would be to disregard the other two sets of innovations which contradict this grouping. Admittedly, such a procedure may be tenable in some cases. For example, C could have undergone some of the same innovations as A and B purely by chance, so that these are not really "shared innovations" in the relevant sense, but are rather "*parallel* innovations". The trouble with this argument is that it is often extremely difficult to come up with positive evidence for it. In particular, if it is believed that C was still in contact with A and B at the time it underwent these innovations, it is unparsimonious to invoke independent, parallel development as an explanation: it is more probable that the changes they have in common reflect events of language-internal diffusion across dialects.

Another situation in which it may be reasonable to disregard the B–C and A–C innovations is when there is good reason to believe that these all occurred historically *after* the A–B innovations, and at a point in time when C had already become mutually unintelligible with A and B (i.e. had become a separate language). In this case, many historical linguists would label the B–C and A–C innovations as effects of "language contact" (or "horizontal transmission"), and would disregard them for the purpose of representing genealogical relationships. This sort of reasoning only works under the assumption that it is possible to draw a principled line between diffusion across language boundaries ("contact") and diffusion within them ("internal change"). This seems unlikely, given that the concept of a "language boundary" (i.e. whether two speech varieties are separate languages or simply dialects of the same language) is itself a gradient notion. However, the argument of contact is usually proposed in good faith, and may be accepted in some obvious cases, namely when the genealogical distance between the speech varieties involved was already much too great at the time of contact for mutual intelligibility—e.g. lexical borrowings from Old Norse into Old English, or from Polynesian languages into other Oceanic languages (Biggs 1965).

In sum, given a set of changes with overlapping distributions, there are occasionally *bona fide* reasons for arguing that some of them are *not* genealogical in nature, and thus should be discarded for the purpose of subgrouping. In general, though, there is often no legitimate basis for deciding which ones may be ignored. Sometimes, this is merely due to lack of evidence (historical or linguistic) about which set of changes predates the other. But in many cases, the problem is simply that the tree model fails to capture the fact that innovations do spread in entangled patterns across sets of mutually intelligible dialects, resulting in intersecting genealogical subgroups. This is what happens in dialect chains and networks, as well as in full-fledged language families that have evolved out of dialect networks—which Ross (1988:8; 1997:213) calls ***linkages***. The relationships among Spanish, French and Italian—or among other Romance languages, for that matter (with the possible exception of Romanian)—are typical of a linkage. Crucially, linkages are common throughout the world: similar configurations have been described—under various names—for Sinitic (Hashimoto 1992; Chappell 2001), Semitic (Huehnergard & Rubin 2011), Indo-Aryan (Toulmin 2009), Athabaskan (Krauss & Golla 1981; Holton 2011), Oceanic (Geraghty 1983, Ross 1988), and many other language families. In Section 4, we will be presenting a detailed example from a section of the Oceanic linkage.

In the case of linkages, decisions about which innovation-defined groupings should be ignored for the purpose of representing genealogical relationships tend to be *ad hoc*, and debates rage with no sign of resolution. In our view, such problems are mere artefacts of the assumptions present in the tree model, and lack any legitimate basis as far as language change is concerned. In fact there is no justification to the assumption that dialects and languages evolve primarily by splitting in a tree-like fashion: the more is known about language change, the more it becomes obvious that this model is a poor approximation of reality, and rests on a misleading metaphor.

In the remainder of this paper, we advance a more flexible model: Historical Glottometry. It elaborates on the principles of the Comparative Method, yet attempts to liberate it from the misleading influence of the family-tree model, by proposing a representation that reflects historical reality more faithfully.

3. Defining Historical Glottometry

3.1. Intersecting subgroups

Insofar as our model is meant to describe (past or present) dialect networks, it is useful to start by looking at how these are represented by dialectologists. A key concept in dialectology is that of the ***isogloss*** (Chambers & Trudgill 1998: 89). Considering a given linguistic property, and the way it is geographically distributed across a dialect network, an isogloss is a line delimiting the set of dialects (or of "communalects", to use a term neutral between "language" and "dialect") that share that property. Isoglosses can be represented on geographically realistic maps, or on more abstract figures. The lines in Figure 4 above are examples of isoglosses, showing the distribution of certain linguistic properties in (part of) the Romance family.

In principle, an isogloss may involve any property that is shared among languages, regardless of its historical origin. And indeed, because dialectology traditionally examines modern speech varieties from a purely synchronic perspective, isogloss maps often fail to distinguish between those similarities that result from shared innovations (*synapomorphies*) and those that are simply shared retentions from a common ancestor (*symplesiomorphies*), or even parallel innovations (*homoplasies*) and accidental similarities.[7] From the perspective of historical linguistics that concerns us here, it is indispensable to restrict our observations to *shared innovations*: indeed, as per Leskien's principle mentioned above, it is a pillar of the Comparative Method that only innovations are indicative of the shared history of communities. The methodology we propose here can be seen as exactly this: a dialectological approach to language history, combining the precise descriptive tools of dialectology and dialectometry (Goebl 2006, Nerbonne 2010, Szmrecsányi 2011) with the powerful concepts of the Comparative Method—notably the stress on shared innovations.

One problem with isogloss maps (and admittedly the main reason why they have not been adopted more widely outside of dialectology) is that they become visually messy very quickly as more and more intersecting isoglosses are added; furthermore, they do not lend themselves to straightforward storytelling as much as a tree diagram would. The former issue, at least, can be addressed if we choose to use isoglosses to represent not individual innovations, but rather language groupings defined by one or more exclusively shared innovations (in other words, subgroups, in our extended sense of the term). A subgroup is simply a grouping of dialects or languages identified by a bundle of (innovation-defined) isoglosses. The thickness of the isogloss line can then be used to represent the strength of the evidence for each language grouping. For example, Figure 7 translates visually the fact that, while the three subgroups AB, AC and BC are all empirically supported, BC is the weakest pairing, and AB the strongest.

[7] Important exceptions include the "dialect map of the Indo-European languages" in Anttila (1989: 305), which is extremely similar in spirit to the model we will be proposing below, as well as the diagrams in Southworth (1964), which are less so. We are grateful to Malcolm Ross for having brought these works to our attention.

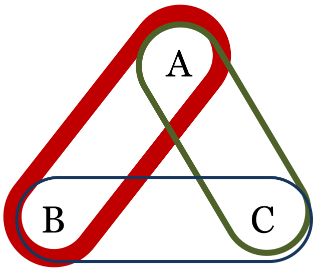


Figure 7: A representation of intersecting subgroups with relative weighting

With such a configuration of the data, historical linguists who take the tree model for granted might be tempted to favour AB as the only valid subgroup, and dismiss the evidence for the two other subgroups altogether, under the assumption that these "weaker" groupings must be mere illusions—whether their similarities be due to "contact", or to "parallel innovation", etc. However, unless there is indeed a principled way of ruling out these isoglosses, it is wiser to keep them in the picture: the idea is that those innovations that are shared between A and C, or B and C, reflect historical events of shared linguistic development just as much as do those between A and B. It is just that the social relations between communities A and B, over the entire course of the history of the ABC family, have been stronger, more frequent or more sustained than those between other pairs of communities. Historical Glottometry can be used precisely as a means to explore and evaluate the strengths of historical connections between social groups, based on the linguistic traces they left in modern languages.

In sum, linguistic linkages make it necessary to accept the idea of a language family in which genealogical subgroups have different strengths, and can cross-cut. Rather than a simplistic binary answer (*X forms vs. does not form a subgroup with Y*), subgrouping studies should allow for the possibility of *stronger vs. weaker subgroups*. Just as a village A may have more frequent mutual interaction with another village B than with C, likewise languages A and B can be said to form a stronger subgroup together (i.e., be "more subgroupy") than languages A and C. Ideally, such claims could even be quantified—as in "A subgroups *n* times as strongly with B as it does with C".[8]

The crucial question is now: how can we define, and calculate, the "strength" of a subgroup? This is the object of the next subsection.

3.2. The cohesiveness of subgroups

The most obvious way to represent the strength of a subgroup using isoglosses would be to simply make the thickness of isoglosses directly proportional to the number of innovations defining the respective groupings. For example, suppose that in the above example of languages A, B and C, there were 12 innovations exclusively shared between A and B, 4 between A and C, and 2 between B and C: then our diagram would look exactly as in Figure 7 (where 1 shared innovation = 1 pixel).

[8] A further extension of our model, which we will not have room to develop in this study, could be to provide both *quantification* and *qualification* to genealogical relations. Thus one could imagine statements along the lines of "A subgroups with B twice as strongly as it does with C as far as regular sound change is concerned; but it does so 1.6 times more with C than with B with respect to verbal morphology, 3 times with respect to lexical replacement in basic vocabulary", etc.

However, suppose that instead of 4 exclusively shared innovations between A and C, there were 24. Our diagram would then be as in Figure 8:

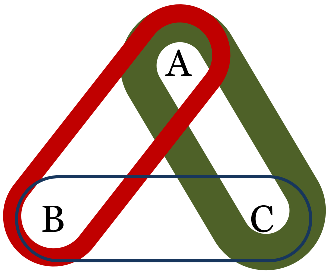


Figure 8: Intersecting isoglosses, with more support for AC than for AB.

Insofar as the thickness of their lines is exactly proportional to the number of exclusively shared innovations between each pair of languages, Figures 7 and 8 are accurate, fully-detailed representations of their respective data. However, they fail to represent an important fact: that the strength of the AB grouping in the first situation is *greater*, relative to the other isoglosses, than the strength of the same grouping in the second situation—*despite* the fact that the same number of defining innovations ($n = 12$) is involved in both cases.

Interestingly, Pawley (2009: 13), discussing the factors that provide evidence for a particular subgrouping hypothesis, notes that "The weight of this evidence depends on the number and quality of the innovations concerned *and on the number and quality of innovations that have conflicting distributions*" (our emphasis). We thus need to quantify the strengths of groupings in a way that takes into account not only the absolute number of innovations that **support** the grouping, but also the number that **conflict** with it. An isogloss x is said to "conflict" with a subgroup y if they cross-cut each other—i.e. if and only if x contains some but not all members of y, and also contains members outside y (mathematically speaking: $x \cap y$, $x \setminus y$ and $y \setminus x$ are all nonempty). In our case, even though the AB grouping is supported by 12 innovations in both cases, it is *more strongly* supported in the first case (where the 12 innovations of AB conflict with only 4 isoglosses for AC plus 2 for BC) than in the second (where the number of conflicting isoglosses is 24 + 2).

In the spirit of procedures common in Social Network Analysis (see Valente 1995, Carrington *et al.* 2005), we propose to define the "**cohesiveness**" of a subgroup as the proportion of **supporting** evidence with respect to the entire set of relevant evidence. Thus, for each given subgroup G, let p be the number of supporting innovations, and q the number of conflicting innovations. The total amount of evidence that is relevant for assessing the cohesiveness of G is $(p + q)$.[9] Now, if we call k_G the cohesiveness value of G, we have:

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

In the situation depicted in Figure 7, the cohesiveness of AB would be calculated as:

[9] Those innovations that are entirely nested within a subgroup (e.g. those that affected only the language B within AB, and no language outside AB) are irrelevant to the cohesiveness of that subgroup, and therefore do not take part in the calculations.

$$k_{AB} = \frac{12}{12+(4+2)} = \frac{12}{18} = \frac{2}{3} \approx \mathbf{67}\%.$$

This result can be translated into plain language by saying that, out of all the innovations that affected the subgroup AB (i.e. either encompassed the subgroup as a whole, *or* affected one of its members together with an external member), exactly two thirds confirmed the cohesion of AB as a subgroup, while one third contradicted it. More simply, A and B “moved together” two-thirds of the time, and “moved apart” one-third of the time.

In the situation depicted in Figure 8, the cohesiveness of AB would be:

$$k_{AB} = \frac{12}{12+(24+2)} = \frac{12}{38} \approx \mathbf{32}\%.$$

That is, in Figure 8, AB as a subgroup is confirmed 32% of the time, and contradicted 68% of the time.

These rates of 67% and 32% should be compared with the theoretical cohesiveness values which all subgroups are supposed to have in a “well-behaved” family tree, namely 100%. In an ideal tree, any group of languages defined by even a single shared innovation is supposed to *always* behave like a subgroup: that is, 100% of the innovations that affect it should confirm its cohesion, and there should be no genealogical innovation involving one (but not all) of its members together with a non-member. As we will see below with real data, this extreme figure of 100% is a convenient fiction that is virtually never met with among real-life languages—at least not in situations of linkages. Rates of cohesiveness in most subgroups typically fall far short of the “ideal” (in our data, most of them have a cohesiveness of between 10% and 30%). This does not mean that we are not dealing with genealogical subgroups at all; but rather, that this very notion must be redefined so as to accommodate the heterodox notion of the *strength* of a subgroup.

3.3. Subgroupiness

Given this measure of cohesiveness, we could use it to determine the thickness of our isogloss lines. However, cohesiveness alone is not sufficient to provide an accurate representation of each subgroup’s strength: as we will see now, it is necessary to also take into account the absolute number of exclusively shared innovations.

Consider now a family of four languages, A, B, C and D, where there are 12 innovations shared by ABC; 4 by AD; 2 by CD, and 1 by AC, as in Figure 9:

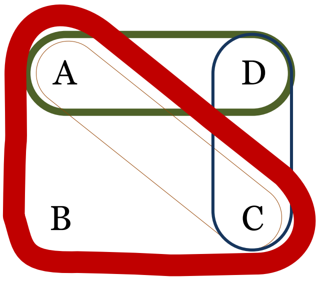


Figure 9: A family of four languages.

Note here that the number of innovations shared by AC ($n = 1$) is irrelevant to the calculation of the cohesiveness of ABC, since it neither confirms this subgroup nor contradicts it (see fn. 11). In order to assess the cohesiveness of ABC, what we need is to observe the number of innovations that confirm it (n = 12) and those that clearly conflict with it—i.e. the innovations of AD (n = 4) plus those of CD (n = 2). The cohesiveness of ABC is thus:

$$k_{ABC} = \frac{12}{12+(4+2)} = \frac{12}{18} = \frac{2}{3} \approx \mathbf{67}\%.$$

Let us now calculate the cohesiveness of AC. This grouping is confirmed not only by the innovations that are exclusively shared by A and C (n = 1), but also by those which they share non-exclusively, since these too show that languages A and C tend to undergo the same linguistic changes together. This includes, in Figure 9, the 12 innovations shared by ABC. As a result, the cohesiveness of the grouping AC should be like this:

$$k_{AC} = \frac{12+1}{(12+1)+(4+2)} = \frac{13}{19} \approx \mathbf{68}\%.$$

In sum, the cohesiveness of AC is even *greater* than that of ABC. Yet we would not want to say that AC is a "stronger" subgroup than ABC, because the latter has a far greater number of *exclusively* shared innovations.

Our proposed solution to this problem is to use the absolute number of *exclusively shared innovations* as the main point of reference, and qualify it using the subgroup's cohesiveness rate (k) as a **weighting** coefficient. For each given subgroup G, let ε be its number of exclusively shared innovations; p its number of supporting innovations (i.e. shared innovations, whether exclusively or not), and q the number of conflicting innovations. We already saw that the *cohesiveness* rate is $k = \frac{p}{(p+q)}$. We now propose to define the **subgroupiness** of a language cluster (call it 'sigma', ς) as the product of its cohesiveness rate (k) with its number of exclusively shared innovations (ε):

$$\varsigma = \varepsilon \times k = \boldsymbol{\varepsilon} \times \frac{\boldsymbol{p}}{(\boldsymbol{p}+\boldsymbol{q})}.$$

For example, if we come back to the comparison of Figures 7 and 8, we can now weight the absolute number of innovations exclusively shared by A and B (ε_{AB}) using AB's cohesiveness rate k_{AB} (given above), and thus calculate its subgroupiness ς_{AB}.

In Figure 7: $\varsigma_{AB} = 12 \times \frac{12}{18} = \mathbf{8}.$

In Figure 8: $\varsigma_{AB} = 12 \times \frac{12}{38} \approx \mathbf{3.79}.$

These numbers constitute exact measurements of the extent to which AB is a more strongly-supported subgroup in the first case than in the second case. (In other words, we can now say, "AB is more than twice as strongly supported—or more simply, *more than twice as subgroupy*—in Figure 7 than in Figure 8".) As for Figure 9, we find that

$$\varsigma_{ABC} = 12 \times \frac{12}{18} = \mathbf{8} \text{ and } \varsigma_{AC} = 1 \times \frac{13}{19} = \frac{13}{19} \approx \mathbf{0.68},$$

in other words, that ABC is more than eleven-and-a-half times as subgroupy as AC. These results are consistent with the intuition that the subgroup ABC is more strongly supported than AC. In conclusion, subgroupiness constitutes the best criterion we have found for assessing the relative strengths of the genealogical subgroups in a language family.

3.3. A visual representation

In terms of visual representation, it is then easy to draw lines around subgroups, whose thickness is proportional to their calculated subgroupiness ς. Figures 7′–9′ show our proposed representations of the situations depicted in Figures 7–9, respectively. We call these kinds of figures **historical glottometric diagrams** ('glottometric diagrams' for short).

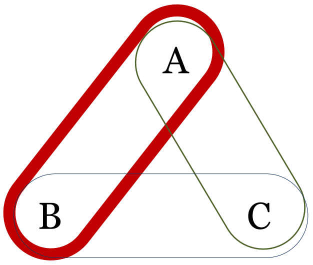


Figure 7′: Illustration of subgroupiness-based isogloss thickness for the situation depicted in Figure 7. Subgroupiness rates: $\varsigma_{AB} = 8$; $\varsigma_{AC} = 0.89$; $\varsigma_{BC} = 0.22$.

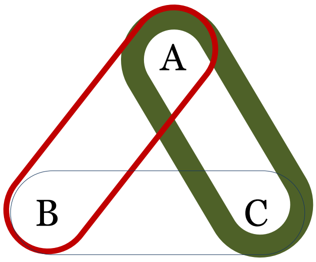


Figure 8′: Illustration of subgroupiness-based isogloss thickness for the situation depicted in Figure 8. Subgroupiness rates: $\varsigma_{AB} = 3.79$; $\varsigma_{AC} = 15.16$; $\varsigma_{BC} = 0.11$.

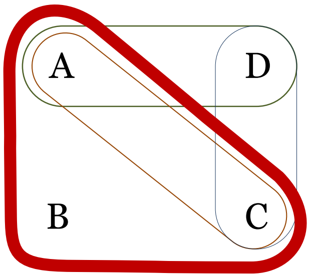


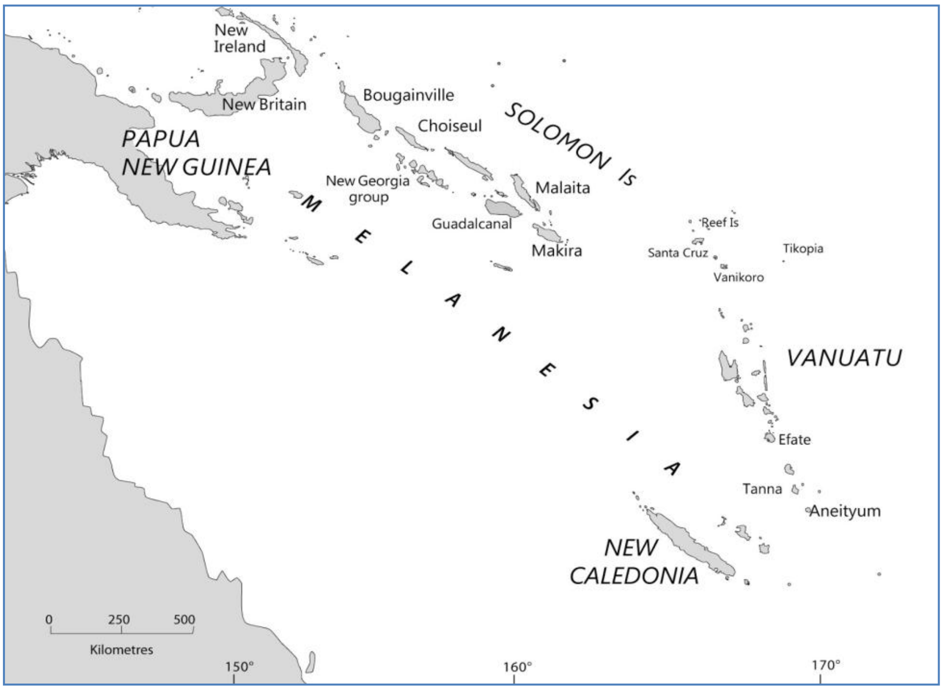
Figure 9′: Illustration of subgroupiness-based isogloss thickness for the situation depicted in Figure 9. Subgroupiness rates: $\varsigma_{ABC} = 8$; $\varsigma_{AD} = 0.84$; $\varsigma_{AC} = 0.68$; $\varsigma_{CD} = 0.21$.

The examples given in this section were abstract, and simple in the sense that they involved small numbers of languages and of innovations. But the same tools can be profitably applied to a much richer set of data. The next section will show precisely how Historical Glottometry can be applied to a real dataset involving 17 languages, and a total of 474 innovations.

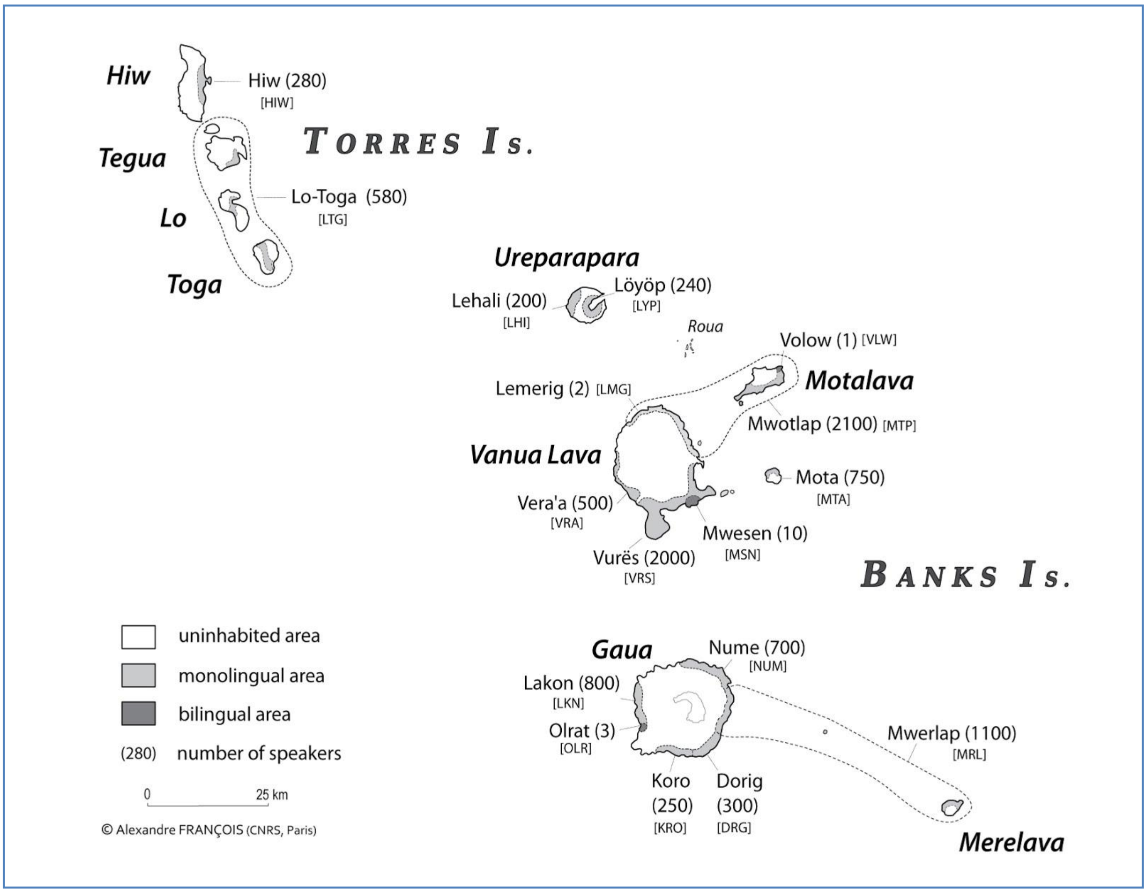
4. A case study from North Vanuatu

4.1. The languages

We can now illustrate the power of Historical Glottometry using a set of actual data from the languages of Vanuatu, an archipelago in the south Pacific (see Map 1).



Map 1: The archipelago of Vanuatu, in the South Pacific



Map 2: The 17 languages of the Torres and Banks Is, in northern Vanuatu

There are around 110 indigenous languages spoken in Vanuatu, which all belong to the Oceanic branch of the Austronesian language family. The evidence for Oceanic being a (classical, nearly 100% cohesive) subgroup of Austronesian is massive (Pawley &

Ross 1995; Ross 1988), and it is widely accepted that there was at some point a more-or-less homogeneous Proto-Oceanic language spoken throughout most of the south Pacific (Pawley & Green 1984; Pawley 2008, 2010), which gradually fragmented into dialects and then independent languages—following a scenario quite similar to the history of Romance languages. Over the decades, there have been a number of attempts to fit the modern-day languages of Vanuatu into a tree model. Clark (2009:4-9) lists as many as nine conflicting subgrouping hypotheses, none of which has reached consensus. This tends to confirm our hypothesis that the genealogical relations among Vanuatu languages cannot be rendered by a tree: they constitute a *linkage*, i.e. a group of modern languages which emerged through the *in situ* diversification of an earlier dialect network (Tryon 1996; François 2011a, 2011b).

We will be focusing on the two northernmost island groups of the Vanuatu archipelago, the Torres and Banks Islands. Alexandre François has been conducting fieldwork there since 1997, and has collected extensive data on the 17 languages still spoken in this small area, many of which are endangered (see François 2012). The names of these languages are given on Map 2, together with three-letter abbreviations and numbers of speakers.

4.2. Intersecting isoglosses in North Vanuatu

The communalects of North Vanuatu have now lost mutual intelligibility, and constitute distinct languages. However, it is possible, thanks to the Comparative Method, to unravel the various linguistic changes that took place since the time of earlier linguistic unity, and brought about the present linguistic diversity (François 2005, 2011a, 2011b). Even though some changes affected a single communalect in isolation, the most typical case was for a given innovation to emerge in some location, and diffuse via social interaction from one dialect to its neighbours, until it settled down into a certain portion of the dialect network. Some isoglosses encompassed the entire area, while others only targeted a set of four or five villages. And of course, in a manner similar to Romance dialects, what we see is that the isoglosses defined by the various innovations cross-cut each other.

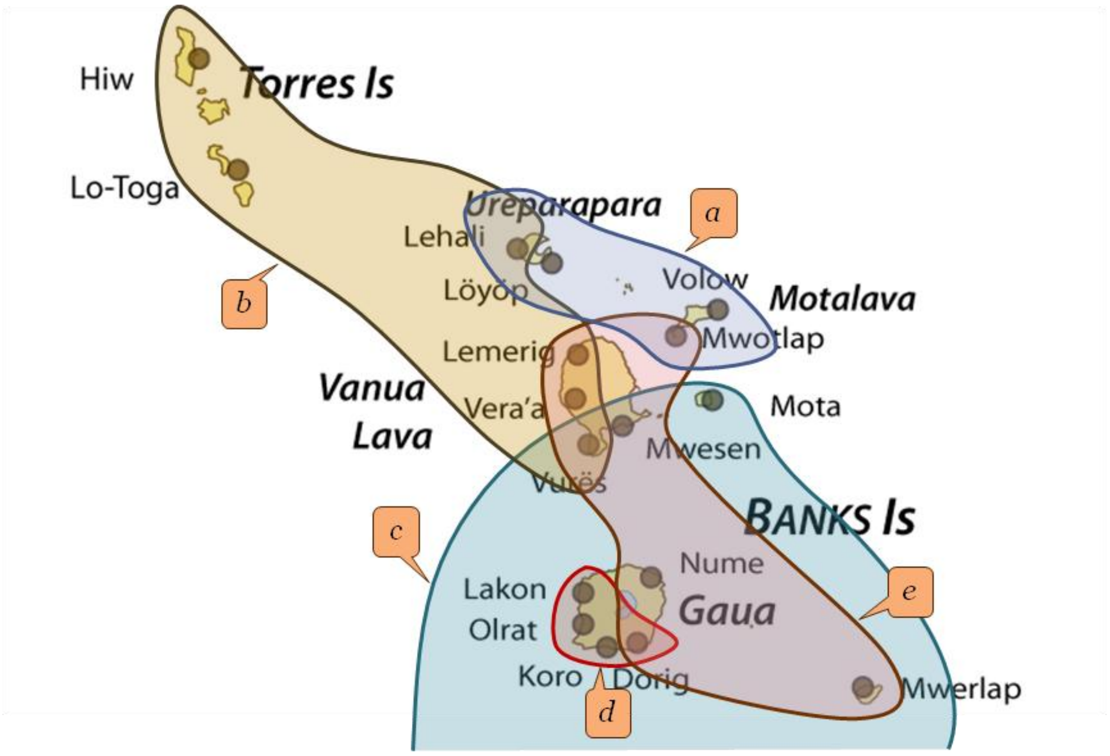
The innovations under discussion here are of various kinds (François 2011a:192–211). They include regular phonological change; irregular sound change (which affects one or a few words rather than applying across the lexicon); morphological change; syntactic change; and lexical replacement. Map 3 shows a selection of isoglosses for the following five innovations:

a) Regular sound change:	**r* > /j/
b) Irregular sound change:	**malate* → **malete* ‘broken’
c) Irregular sound change:	**ʔaŋaRi* → **ʔaŋai* ‘almond’
d) Morphological change:	metathesis in trial pronouns (Plural+three → three+Plural)
e) Morphological change:	**toɣa* ‘stay’ → Prohibitive

Map 3 makes it clear that isoglosses in the Torres and Banks languages—like those in the Romance family—constantly intersect.[10] There is no way the genealogical relations among these languages could be represented by a tree. François (2004) was an attempt to do precisely this; while a tentative tree was indeed proposed, the number of

[10] Note that one innovation, namely (c), involves not only a subset of the Banks languages, but also languages further south in Vanuatu (François 2011b:157).

issues raised (conflicting evidence, intersecting isoglosses, the need to constantly resort to *ad hoc* hypotheses to preserve the tree structure) were preliminary signs of the inadequacy of the cladistic approach in this part of the world.



Map 3: Five isoglosses in the Torres–Banks Islands

What we need here is a Historical Glottometry approach, which will tell us, amongst the 131,070 (= $2^{17} - 2$) potential groupings involving these languages, which ones actually exist, and constitute the strongest subgroups. That these subgroups will probably intersect is to be expected, and is no longer a problem: by now there is good reason to believe that this is the default situation in most language families. What we need is simply to go beyond the observation of individual isoglosses as in Map 3, and be able to base our calculations on a rich database.

4.3. Identifying innovations

4.3.1. Applying the Comparative Method

Our dataset consists of a table of 474 separate innovations which A. François identified in these 17 languages. For each linguistic feature considered, systematic comparison was conducted among languages of the sample as well as with other Oceanic languages, following principles of the Comparative Method, so as to establish the ancestral state of each property in the languages' shared ancestor (Proto-Oceanic, or a close variant thereof) as well as the direction of change.

Some cases make it relatively easy to determine what the innovation was. For example, consider the words for 'almond': whereas the eight languages to the north reflect the protoform ***ʔaŋaRi* (e.g. Vera'a *ŋar*), the languages further south reflect a form ***ʔaŋai* (e.g. Vurës *ŋɛ*). The latter protoform shows the irregular loss of *R, a frequent yet lexically-specific sound change in the area (François 2011b). It is clearly an innovative form, whose distribution in the Banks Islands is represented by isogloss (c) in Map 3.

In other cases, identifying the innovation requires more reflection. For example, most of the northern Vanuatu languages have an adjective meaning ‘broken’, with forms that are cognate with each other:

(1) ‘broken’: HIW *mɪjɪt*; LTG *məlit*; LHI *mɛlɛt*; LYP *malat*; VLW *malat*; MTP *malat*; LMG *mɛlɛʔ*; VRA *mɪlɪʔ*; VRS *mɪlɪt*; MSN *malat*; MTA *malate*; NUM *malat*; DRG *mlat*; MRL *mɛlɛt*.

One can show that these modern forms go back to two distinct protoforms: **malate* and **malete*. This conclusion is based on our knowledge of regular sound changes in this area, established using the Comparative Method (François 2005). This allows us to discern even those cases where two cross-linguistic homophones derive from different etyma: for example, while Lehali /mɛlɛt/ necessarily reflects **malete*, the same surface form /mɛlɛt/ in Mwerlap is a regular reflex of **malate*, because a stressed /a/ followed by an unstressed /e/ in the next syllable regularly underwent umlaut in this language (*aCe > /ɛC/). Knowledge of each language’s phonological history likewise enables us to link each modern form in (1) to one, and only one, of the two protoforms—either **malate* or **malete*. The next, crucial step consists in determining which of these two is conservative, and which one is innovative. External evidence is indispensable here, and shows that other Oceanic languages outside the Torres–Banks area point to the form with /a/: e.g. Araki /n̼alare/ ‘broken’ < **malate* (François 2002:270). In sum, the innovation we are concerned with here is a lexically-specific, irregular sound change whereby **malate* became **malete*, and not the other way around. The languages that participated in this particular innovation are: Hiw, Lo-Toga, Lehali, Lemerig, Vera’a and Vurës. This innovation is represented with isogloss (b) in Map 3.

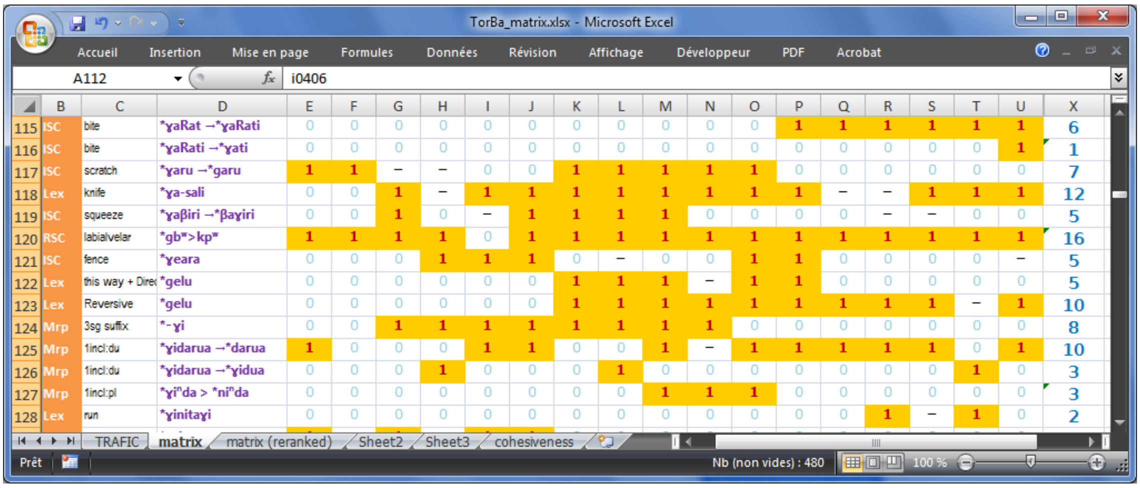
4.3.2. Creating the dataset

The sort of reasoning illustrated above, which follows a rigorous application of the Comparative Method, was used to identify all 474 innovations. The distribution of innovations into various types was as follows:

NATURE OF CHANGE	NUMBER	PROPORTION
Regular sound change	21	4 %
Irregular sound change	116	25 %
Morphological change	91	19 %
Syntactic change	10	2 %
Lexical replacement	236	50 %
Total	474	100 %

Among these types of changes, we consider irregular sound change and morphological change to be the most diagnostic of historical relatedness (following Greenberg 1957:51, Ross 1988:12), because they are least likely to be independently innovated. Lexical material is often excluded from subgrouping studies under the assumption that it is easily borrowable; to avoid this (perceived) problem, we have included here only those lexical replacements which can be shown to predate events of (regular or irregular) sound change.[11]

Figure 10 shows what the final database looks like. The 17 languages are ranked from north-west to south-east; each row corresponds to one innovation, and indicates whether there is positive evidence that a language participated (1) or did not participate (0) in the innovation. An empty box (–) was used when the data are inconclusive, non-applicable, or simply lacking. Altogether, the database contains 2728 positive ('1'), 5040 negative ('0') and 290 agnostic ('–') data points.



	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	X
115	ISC	bite	*ɣaRat →*ɣaRati	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	6
116	ISC	bite	*ɣaRati →*ɣati	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
117	ISC	scratch	*ɣaru →*garu	1	1	–	–	0	0	1	1	1	1	1	0	0	0	0	0	0	7
118	Lex	knife	*ɣa-sali	0	0	1	–	1	1	1	1	1	1	1	1	–	–	1	1	1	12
119	ISC	squeeze	*ɣaβiri →*βaɣiri	0	0	1	0	–	1	1	1	1	0	0	0	0	–	–	0	0	5
120	RSC	labialvelar	*gbʷ>kpʷ	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	16
121	ISC	fence	*ɣeara	0	0	0	1	1	1	0	–	0	0	1	1	0	0	0	0	–	5
122	Lex	this way + Direc	*gelu	0	0	0	0	0	0	1	1	1	–	1	1	0	0	0	0	0	5
123	Lex	Reversive	*gelu	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	–	1	10
124	Mrp	3sg suffix	*-ɣi	0	0	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	8
125	Mrp	1incl:du	*ɣidarua →*darua	1	0	0	0	1	1	0	0	1	–	1	1	1	1	1	0	1	10
126	Mrp	1incl:du	*ɣidarua →*ɣidua	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	1	0	3
127	Mrp	1incl:pl	*ɣiⁿda > *niⁿda	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	3
128	Lex	run	*ɣinitaɣi	0	0	0	0	0	0	0	0	0	0	0	0	0	1	–	1	0	2

Figure 9: A sample of our database of historical innovations in the Torres–Banks languages.

Note that each pattern of 1s and 0s corresponds to a diffusion area, and would be represented with an isogloss. We will now illustrate the application of Historical Glottometry to this database, following the methods explained in the previous section.

4.4. The results

4.4.1. Numerical results

The first thing we can do with this dataset is to measure cohesiveness for clusters of two languages. This measure of "pairwise cohesiveness",[12] applied to all pairs of languages ($17^2 = 289$), yields the results in Table 1.

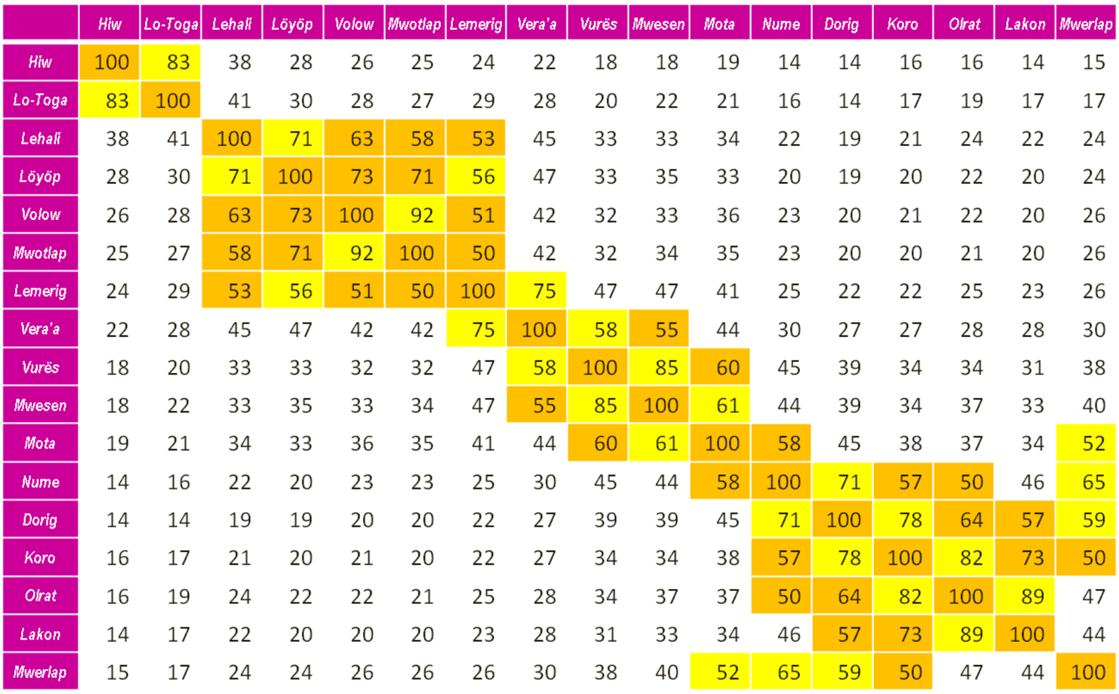
The figures of 100% in the diagonal simply say, as it were, that a language always subgroups perfectly with itself; these can thus be disregarded. More instructive is the observation that the cohesiveness *k* of language pairs tends to vary a lot, but with the highest figure being only 92%. The coloured (yellow and orange) cells indicate rates of 50% and above, i.e. pairs with relatively high cohesiveness.

To illustrate the proper interpretation of the table, the figure of 92%, between Volow and Mwotlap, indicates that when either of these languages underwent a change (to-

[11] This is the same reasoning that validates **manducāre* 'eat' as a legitimate example of early lexical innovation shared by French and Italian (§2.1), because it reflects regular sound changes diagnostic of inherited vocabulary (compare French *manger* /mɑ̃ʒe/ <**manducāre* with *venger* /vɑ̃ʒe/ 'avenge' <**vindicāre*). By contrast, a recent Italian loanword such as *caporal* ('corporal'), which does not exhibit any such sound changes, would not normally qualify as diagnostic evidence for subgrouping.

[12] This is quite similar to the concept of "Relative Identity Weight" in the Salzburg school of dialectometry (Goebl 2006: 412).

gether with some other language), it shared it with the other member of the pair 92% of the time. Table 1 thus shows that languages share innovations with their immediate neighbours a lot of the time—yet they do so at varying rates.



	Hiw	Lo-Toga	Lehali	Löyöp	Volow	Mwotlap	Lemerig	Vera'a	Vurës	Mwesen	Mota	Nume	Dorig	Koro	Olrat	Lakon	Mwerlap
Hiw	100	83	38	28	26	25	24	22	18	18	19	14	14	16	16	14	15
Lo-Toga	83	100	41	30	28	27	29	28	20	22	21	16	14	17	19	17	17
Lehali	38	41	100	71	63	58	53	45	33	33	34	22	19	21	24	22	24
Löyöp	28	30	71	100	73	71	56	47	33	35	33	20	19	20	22	20	24
Volow	26	28	63	73	100	92	51	42	32	33	36	23	20	21	22	20	26
Mwotlap	25	27	58	71	92	100	50	42	32	34	35	23	20	20	21	20	26
Lemerig	24	29	53	56	51	50	100	75	47	47	41	25	22	22	25	23	26
Vera'a	22	28	45	47	42	42	75	100	58	55	44	30	27	27	28	28	30
Vurës	18	20	33	33	32	32	47	58	100	85	60	45	39	34	34	31	38
Mwesen	18	22	33	35	33	34	47	55	85	100	61	44	39	34	37	33	40
Mota	19	21	34	33	36	35	41	44	60	61	100	58	45	38	37	34	52
Nume	14	16	22	20	23	23	25	30	45	44	58	100	71	57	50	46	65
Dorig	14	14	19	19	20	20	22	27	39	39	45	71	100	78	64	57	59
Koro	16	17	21	20	21	20	22	27	34	34	38	57	78	100	82	73	50
Olrat	16	19	24	22	22	21	25	28	34	37	37	50	64	82	100	89	47
Lakon	14	17	22	20	20	20	23	28	31	33	34	46	57	73	89	100	44
Mwerlap	15	17	24	24	26	26	26	30	38	40	52	65	59	50	47	44	100

Table 1: Pairwise cohesiveness values (percentages) among the 17 Torres–Banks languages

These figures, incidentally, are a valuable result in themselves, as they provide an empirical measurement of how much two languages have evolved together throughout their history. For example, the fact that Lo-Toga (#2) and Lehali (#3) shared only 41% of their innovations together points to a rather strong social divide between the Torres islands on the one hand, and the Banks islands on the other hand: clearly, the Lo-Toga community has had much less social interaction with Lehali (k = 41%) than with Hiw (k = 83%). Likewise, it is instructive to observe that, even though the language Vurës is geographically spoken only a couple of hours' walk from Vera'a (see Map 2), the two languages share together no more than 58% of their innovations; the historical links were much stronger, on the one hand, between Vera'a and Lemerig (k = 75%), and on the other hand, between Vurës and Mwesen (k = 85%). Interestingly, these figures closely match the intuitive feel one gets when learning and comparing the languages of Vanua Lava, as well as the islanders' own impressions; except that the figures have the advantage of being precise, and directly comparable with one another.

In order to deserve the status of genealogical subgroup, a cluster of languages needs to be "attested" historically, i.e. have at least one exclusively shared innovation ($\varepsilon \geq 1$). A subgroup uniting Volow and Löyöp, for example, would have high cohesiveness (73%) if it existed; but because no innovation happens to be shared exclusively by these two languages, they cannot count together as a subgroup. Pairings that are not supported by at least one isogloss appear here in orange. Conversely, the yellow cells in Table 1 correspond to those higher-cohesiveness pairings ($k \geq$ 50%) which are actually attested as subgroups: e.g. Hiw–Lo-Toga with 83%, Lehali–Löyöp with 71%, etc.

We applied the same method to calculate the cohesiveness (k) of all attested clusters of North Vanuatu, of any size. In total, the number of unique innovation-defined subgroups was 143. This figure includes the 15 pairs of languages shown in yellow in Table 1 above, but also clusters of various sizes, up to 15 members. The results, which cannot all be presented here for lack of space, were useful for the next stage: the calculation of subgroupiness values (ς).

4.4.2. A glottometric diagram

We calculated the subgroupiness of all 143 attested language clusters, by applying the principles exposed in §3 above. The 15 subgroups with the highest subgroupiness values are listed in Table 2.

Table 2: The 15 strongest subgroups in the Torres–Banks linkage.

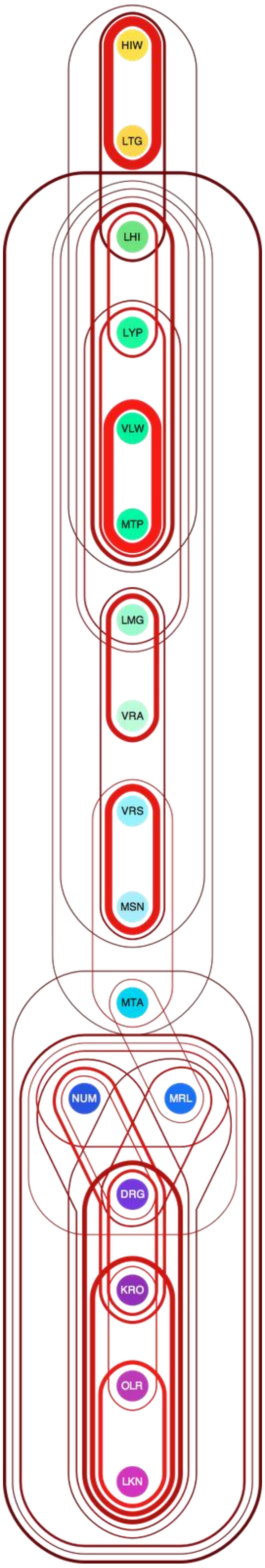
subgroups	*subgroupiness*
Volow–Mwotlap	12.82
Hiw–LoToga	12.45
Vurës–Mwesen	9.34
Lemerig–Vera'a	6.78
Koro–Olrat–Lakon	6.63
Dorig–Koro–Olrat–Lakon	6.01
Olrat–Lakon	5.34
Lehali–Löyöp–Mwotlap–Volow	5.22
15 Banks languages (LHI→LKN)	3.92
Dorig–Koro	3.90
Löyöp–Volow–Mwotlap	3.64
Lehali–Löyöp	3.53
Hiw–LoToga–Lehali	3.43
southern Banks (Mwerlap + Gaua)	2.99
Dorig–Mwerlap	2.37

In terms of visual representation, the abundance of subgroups of varying strengths made it necessary to represent only the strongest ones—we chose to show only those whose subgroupiness value is greater than or equal to 1 (ς ≥ 1). This includes the 15 subgroups listed in Table 2, plus 17 others. We then represented each subgroup's strength by having line thickness proportional to its subgroupiness. In addition, the degree of redness (brightness value of the contour line) was made proportional to its *cohesiveness*, with more cohesive subgroups appearing more intensely red. The final result was a comprehensive **glottometric diagram** of the whole region (Figure 11).

This result would warrant more commentary than is possible in this paper;[13] we will stick to the essentials. First of all, the subgroupiness values, as well as the map derived from them, confirm the statement in §4.2, that the languages of northern Vanuatu form a *linkage* in which isoglosses, and hence subgroups, constantly intersect. Lehali (LHI), for example, subgroups both with the two Torres languages to its north (ς = 3.43) *and* with the other Banks languages to its south (ς = 3.92). 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). No family tree could ever account for this situation.

[13] The colors of the dots representing the languages are also significant, but explaining this would be a task best left to a lengthier paper (François & Kalyan forthc.).

Figure 11: A glottometric diagram of the Torres–Banks languages



It is worthy of notice that the glottometric approach can *also* detect and represent those situations which are "tree-like": for example, Volow and Mwotlap form a subgroup clearly separated from Löyöp; Vurës and Mwesen also clearly belong together. But evidently, these tree-like patches are a rarity in a language network which is strongly non-tree-like.

Another important result is the observation that Torres–Banks languages generally pattern in a geographically coherent way: all languages adjacent on the glottometric diagram are also adjacent geographically (though not vice versa; see below). This is even true for the non-linear part of the map, involving the four languages Mota–Nume–Dorig–Mwerlap: all the language pairs attested there (MTA–MRL, NUM–MRL, NUM–DRG, MRL–DRG) correspond to adjacent languages on Map 2. It is impossible to capture such tight geographical organisation using a tree: any binary tree of 17 languages will allow 65,536 (= 2^{16}) possible linear orderings of languages.

Expected though it may be, this consistency between language history and geography is a valuable result: for it shows that the languages' anchoring in space must have remained stable over the three millennia of their historical development, with limited interisland migration (François 2011b: 181). Applying Glottometry to historically more turbulent families would make it possible to detect the genealogical relations that hold between languages *in spite* of their geographic locations, as accurately as the Comparative Method on which it is based.

And indeed, a finer grain of observation reveals certain non-trivial patterns in our data that do more than just index geography. For example, even though Volow's location is closer to Mota than to Löyöp (Map 2), the position of the three languages in the diagram shows that Volow and Mota are genealogically quite remote (k = 36%). Evidently, the ancient societies of Motalava and Mota islands had very few direct social interactions with each other, and much more with the other islands (Ureparapara, Vanua Lava) located to their west. Such a result illustrates the potential of the method to reconstruct the shape of past social networks.

5. Conclusion

In conclusion, our newly proposed method of Historical Glottometry allows us to escape the false dichotomies of the tree model—e.g. whether it is A–B, A–C, or B–C that is truly diagnostic of genealogical relatedness—by allowing us to posit intersecting subgroups, and to quantify the *strength* of the genealogical evidence in favour of each language cluster.

If we were to use a tree to represent our data, we would certainly be able to capture certain salient organising features, e.g. the split between the two Torres languages (Hiw and Lo-Toga) and all the languages to the south. But a tree would only be able to provide a very distorted picture of the social history of the region—as an orderly sequence of migrations with loss of contact—while the story told by the data (made visible to us by the glottometric diagram) is a much richer and more varied narrative of social interaction in which languages converge as much as they diverge. Far from the approximations imposed by the assumptions of the tree model, we hope to have shown the way towards a more accurate and realistic representation, which stays true to the most valuable insights of the Comparative Method.

References

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

Alkire, Ti & Rosen, Carol. (2010). *Romance Languages: A Historical Introduction*. Cambridge: Cambridge University Press.

Anttila, Raimo. (1989). *Historical and Comparative Linguistics*. (2nd edition). Amsterdam: John Benjamins.

Avise, John C. (2000). *Phylogeography: The history and formation of species*. Cambridge, MA: Harvard University Press.

Berger, Roger & Annette Brasseur. (2004). *Les Séquences de sainte Eulalie*. Geneva: Droz.

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

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

Bossong, Georg. (2009). Divergence, convergence, contact. Challenges for the genealogical classification of languages. In *Convergence and divergence in language contact situations*, ed. by K. Braunmüller & J. House. Hamburg Studies on Multilingualism, 8. Amsterdam-Philadelphia: John Benjamins. Pp.13-40.

Bryant, David, Flavia Filimon and Russell D. Gray. (2005). Untangling our past: Languages, Trees, Splits and Networks. In *The Evolution of Cultural Diversity: Phylogenetic Approaches*, ed. by R. Mace, C. Holden & S. Shennan. London: UCL Press. pp. 69-85.

Campbell, Lyle. (2004). *Historical linguistics: an introduction*. MIT Press.

Carrington, Peter J., John Scott and Stanley Wasserman (eds.). (2005). *Models and Methods in Social Network Analysis*. New York: Cambridge University Press.

Chambers, Jack K., & Peter Trudgill. (1998). *Dialectology*. Cambridge textbooks in linguistics. Cambridge: Cambridge University Press.

Chappell, Hilary. (2001). Language contact and areal diffusion in Sinitic languages. In Aikhenvald, A. and R.M.W. Dixon (eds.), pp. 328–357.

Croft, William. (2000). *Explaining language change: An evolutionary approach*. Pearson Education.

Crowley, Terry & Claire Bowern. (2010). *An Introduction to Historical Linguistics*. (4th edition). Oxford: Oxford University Press.

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

Enfield, Nicholas. (2008). Transmission Biases in Linguistic Epidemiology. In R. Nicolaï & B. Comrie (eds), *Language Contact and the Dynamics of Language: Theory and Implications*. Thema 2. *Journal of Language Contact*, 299-310.

François, Alexandre. (2002). *Araki. A disappearing language of Vanuatu*. Pacific Linguistics, 522. Canberra: Australian National University.

François, Alexandre. (2004). Subgrouping hypotheses in North Vanuatu. Paper presented at the Sixth *International Congress on Oceanic Linguistics* (COOL6). University of the South Pacific, Port Vila, Vanuatu.

François, Alexandre. (2005). Unraveling the history of the vowels of seventeen northern Vanuatu languages. *Oceanic Linguistics* 44 (2):443-504.

François, Alexandre. (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.

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

François, Alexandre. (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.

François, Alexandre. (forthc.). Models of language diversification. In Claire Bowern & Bethwyn Evans (eds), *The Routledge Handbook of Historical Linguistics*. New York: Routledge.

Garrett, Andrew. (2006). Convergence in the Formation of Indo-European Subgroups: Phylogeny and Chronology. *Phylogenetic methods and the prehistory of languages*, ed. by Peter Forster & Colin Renfrew, 139–151. Cambridge: McDonald Institute for Archaeological Research.

Geraghty, Paul A. (1983). *The History of the Fijian languages*. Oceanic Linguistics Special Publication, 19. Honolulu: University of Hawaii Press.

Goebl, Hans. (2006). Recent Advances in Salzburg Dialectometry. *Literary and Linguistic Computing* 21 (4): 411-435.

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

Greenberg, Joseph H. (1957). *Essays in Linguistics*. Chicago: University of Chicago Press.

Hall, Robert A. Jr. (1950). The Reconstruction of Proto-Romance. *Language* 26 (1):6-27.

Hashimoto, Mantarō J. (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.

Heggarty, Paul, Warren Maguire & 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*. (2nd edition). Berlin: Mouton 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, & Aaron Rubin. (2011). Phyla and Waves: Models of Classification. Semitic Languages. In *Semitic Languages: An International Handbook*, edited by S. Weninger, G. Khan, M. P. Streck & J. Watson. Handbücher zur Sprach- und Kommunikationswissenschaft, 36. Berlin: de Gruyter Mouton. Pp.259-278.

Kortlandt, Frederik. (2007). *Italo-Celtic origins and prehistoric development of the Irish language*. Amsterdam: Rodopi.

Krauss, Michael E., & Victor Golla. (1981). Northern Athapaskan languages. In *Handbook of North American Indians*, vol. 6: *Subarctic*, edited by J. Helm. Pp.67-85.

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

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

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

Minaka, Nobuhiro & Kunihiko Sugiyama. (2012). *Keitōju Mandara: Chein/tsurī/nettowāku* [Phylogeny Mandala: Chain, tree, and network] (in Japanese). Tokyo: NTT Publishing.

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

Page, Roderick D. M. & 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. *Selected Papers from the Eighth International*

Conference on Austronesian Linguistics, ed. by Elizabeth Zeitoun & Paul Jen-Kuei Li, 95–138. Symposium Series of the Institute of Linguistics: Academia Sinica.

Pawley, Andrew. (2008). Where and when was Proto-Oceanic spoken? Linguistic and archaeological evidence. In *Language and text in the Austronesian world: studies in honour of Ülo Sirk*, edited by Y. A. Lander & A. K. Ogloblin. München: Lincom Europa. Pp.47-71.

Pawley, Andrew. (2009). Polynesian paradoxes: Subgroups, wave models and the dialect geography of Proto Polynesian. Unpublished paper delivered at the Eleventh *International Conference on Austronesian Linguistics*, Aussois (France), June 2009.

Pawley, Andrew. (2010). Prehistoric Migration and Colonisation Processes in Oceania: A View from Historical Linguistics and Archaeology. In *Migration history in world history: multidisciplinary approaches*, edited by J. Lucassen. Studies in global social history, 3. Leiden: Brill. Pp.77-112.

Pawley, Andrew, & Roger C. Green. (1984). The Proto-Oceanic Language Community. *Journal of Pacific History* 19:123-146.

Pawley, Andrew, & Malcolm Ross. (1995). The prehistory of Oceanic languages: a current view. In *The Austronesians: Historical and Comparative Perspectives*, edited by P.S. Bellwood, J.J. Fox & D. Tryon, Comparative Austronesian Project. Canberra: Australian National University. Pp.39-80.

Posner, Rebecca. (1996). *The Romance Languages*. Cambridge: Cambridge University Press.

Ross, Malcolm. (1988). *Proto-Oceanic and the languages of Western Melanesia*. Canberra: ANU Press.

Ross, Malcolm. (1997). Social networks and kinds of speech-community event. In *Archaeology and language 1: Theoretical and methodological orientations*, edited by R. Blench & M. Spriggs. London: Routledge. Pp.209-261.

Ross, Malcolm, Andrew Pawley, & Meredith Osmond, eds. (2011). *The lexicon of Proto Oceanic: Animals*. Pacific Linguistics, 621. Canberra: Australian National University.

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

Schmidt, Karl-Horst. (1993). Insular Celtic: P and Q Celtic. In *The Celtic languages*, edited by M. J. Ball & J. Fife. New York: Routledge. Pp.64-99.

Séguy, Jean. (1973). La dialectométrie dans l'Atlas linguistique de la Gascogne. *Revue de linguistique romane* 145-146:1-24.

Skelton, P., Smith, A., & Monks, N. (2002). *Cladistics: a practical primer on CD-ROM*. Cambridge University Press.

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

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

Toulmin, Matthew. 2009. *From Linguistic to Sociolinguistic Reconstruction: the Kamta historical subgroup of Indo-Aryan*. Pacific Linguistics: Australian National University.

Tryon, Darrell. 1996. Dialect chaining and the use of geographical space. In *Arts of Vanuatu*, edited by J. Bonnemaison, K. Huffman, C. Kaufmann & D. Tryon. Bathurst: Crawford House Press. Pp.170-181.

Valente, Thomas W. (1995). *Network Models of the Diffusion of Innovations*. Cresskill, NJ: Hampton Press.

Wüest, Jakob. (1994). La restructuration du système des démonstratifs en protoroman, in Jacqueline Cerquiglini-Toulet and Olivier Collet (eds.), *Mélanges de philologie et de littérature médiévales offerts à Michel Burger*, 41–49. Genève: Droz.