SPECIAL ISSUE

Understanding language genealogy
Alternatives to the tree model

Edited by Siva kalyan, Alexandre François and Harald Hammarström

Australian National University / LaTTiCe, CNRS, École Normale Supérieure, Univ. Paris 3 Sorbonne nouvelle, Australian National University / Uppsala University

Table of contents

When the waves meet the trees

A response to Jacques and List

Siva Kalyan[1] and Alexandre François[1,2]
[1] Australian National University
[2] LaTTiCe (CNRS; ENS-PSL; Paris 3-USPC)

1. Introduction

We thank Jacques and List (henceforth, J&L) for their paper in defense of the family tree model in historical linguistics. They present powerful arguments in favor of the view that the tree model should be taken as the default model of language genealogy – except in cases of hybridization, which are rare.[1]

We find J&L's use of the notion of "incomplete lineage sorting" (Section 4.1) to be illuminating, and have learned much from their discussion of "undetectable borrowings" and loanword nativization (Section 4.2). However, we believe they have misunderstood the aims of Historical Glottometry (François 2014, 2017; Kalyan & François 2018), the model of language diversification that it assumes, and our reasons for making certain methodological choices when applying it. By clarifying these points, we hope to show that our approach to language diversification – and that of other researchers who subscribe to a wave-based approach to language genealogy – is in fact largely compatible with that of J&L.

2. Two ways of understanding the family-tree model

We believe that there is considerable difference between J&L's understanding of the family tree model and the way in which it is traditionally understood and applied in historical linguistics. Historical Glottometry (and other instantiations of the wave model) arose as a critique of historical linguists', not phylogeneticists', use of the tree model; thus, in order to understand the aims of Historical Glottometry and evaluate how well it fulfills these aims, it is necessary to clarify how historical linguists traditionally use and understand the tree model.

1. See, however, Nichols (1992) and Donohue (2013: 222–224) on the prevalence of areal diffusion of structural features.

https://doi.org/10.1075/jhl.18019.kal
Journal of Historical Linguistics 9:1 (2019), pp. 167–176. issn 2210-2116 | e-issn 2210-2124

In the traditional understanding (see François 2014: 163–167 for more detailed discussion), a population speaking an internally-homogeneous language separates into two or more sub-populations. Each sub-population undergoes linguistic changes, then splits into sub-populations of its own, and so on recursively through time. Crucially, as long as a language community remains in existence – i.e., until it splits into two separate sub-populations – the traditional tree model assumes that every (viable) language change must always have spread to the *entire* population; in biological terms, every linguistic innovation is assumed to have "gone to fixation" before any further event of lineage splitting.[2] As a consequence, every innovation is assumed to be passed on to all of the descendants of the language in which it occurs – and *only* to those descendants. This leads to the useful result known as "Leskien's principle" (after Leskien 1876: vii, though the principle was first clearly stated by Brugmann 1884: 231): namely, that a subgroup of a language family – a set of languages that descend from a single intermediate proto-language – can only be identified by verifying whether those languages are defined by a set of *exclusively-shared innovations*. This principle of subgrouping by exclusively-shared innovations is an essential part of the Comparative Method underlying the practice of historical linguistics, and is also reflected in the assumptions of Bayesian gain-loss phylogenetic methods (see Greenhill & Gray 2012: 525–526).

However, it is often the case (both for languages and for species: see Baum & Smith 2013: 146–151) that an innovation only spreads partway through a population before that population splits. In this situation, an innovation need not be passed on to *all* of the descendants of the language it occurs in, but only to some of them. Moreover, since different segments of the original population may have undergone different partially-diffused innovations, different innovations may be passed on to different subsets of descendants. This is the fundamental observation of the wave model: that innovations within a language (i.e. among closely related dialects) typically show overlapping patterns (Bloomfield 1933: 317; de Saussure 1995 [1916]: 273–278).

There are two ways to apply the insights of the wave model to the subgrouping of a language family:

– The first is to define a "subgroup" as "all of the languages descended from a single intermediate proto-language" and accept that some subgroups may not be definable in terms of exclusively-shared innovations, but only in terms of a chain or network of overlapping innovations.
– The second is to adhere to the definition of "subgroup" used in the traditional application of the tree model – namely, as a set of languages defined by

2. As noted by Baum & Smith (2013: 79), this is often a reasonable assumption in biology, as the rate of fixation of novel traits is extremely rapid compared to the rate of lineage splitting.

exclusively-shared innovations – and accept that subgroups (thus defined) may intersect rather than being strictly nested. This is the approach that we take in Kalyan & François (2018), where we explicitly reconcile the wave model with the Comparative Method.

We tend to prefer the latter, more easily operationalized definition of "subgroup," while J&L adhere to the former one (as does, e.g., Ross 1997 in his discussion of "innovation-defined" vs. "innovation-linked" subgroups). We see value in both approaches.

In both the traditional understanding of the tree model and the more nuanced understanding outlined above, it is taken for granted that languages diversify as a result of successive population splits. The main difference, then, is whether each population split necessarily leads to the emergence of a homogeneous daughter language (as assumed in the traditional family-tree model: see Ross 1997: 212) or whether the daughter language may exhibit dialectal variation from the very beginning due to arising from a segment of a dialect network. Thus, we do not contest J&L's statement that our approach "silently acknowledges … tree-like divergence … even if it turns out to be a star-phylogeny"; in fact, we wholeheartedly acknowledge it.

It is in fact possible to restate our views in terms of the framework proposed by J&L. If we allow for the partial diffusion of an innovation through a proto-language, then we are effectively allowing for variation in the proto-language that is differentially resolved in each descendant. This is nothing more than *incomplete lineage sorting*, but of a kind that J&L do not discuss: namely, one in which the variation in the proto-language is conditioned areally rather than morphologically or socially. This term introduced by J&L provides historical linguists with a useful, biologically-inspired way of talking about the phenomena highlighted by the wave model.

3. Methodology and goals of Historical Glottometry

In the remainder of this response, we address J&L's specific remarks on Historical Glottometry (in their Section 3.3), and attempt to clarify our position where we feel it has been misunderstood.

3.1 Definition and identification of "shared innovations"

J&L criticize our use of the term "shared innovation" for two reasons. Firstly, they object to our using the term agnostically for both "true" shared innovations as well as innovations that might turn out to be cases of parallel development, and suggest that our dataset should consist of only those innovations that are securely known

to belong to the first category. Secondly, they claim that if we were to restrict our dataset in this way, we would not find any overlapping innovations: that shared innovations in a tree can never overlap, because each must occur in a single node.

We believe that it is necessary for us to be agnostic at the data-assembly stage about whether the innovations we identify are shared or parallel; the very purpose of the glottometric method is to *infer* which sets of innovations are likely to be "true" shared innovations (by virtue of following a consistent pattern) and which are likely to be parallel innovations (by virtue of following a pattern that is infrequently attested).[3] In other words, the assumption we make is that cases of parallel innovation will "come out in the wash" – and indeed, in our work, we have found that innovations that follow a geographically haphazard pattern (and are thus likely to be cases of parallel innovation) are invariably associated with subgroups that have low rates of cohesiveness and subgroupiness.

We agree that each (genealogically-relevant) innovation must occur at a single node in the tree – i.e., in a single ancestral speech community; however, this does not mean that innovations within this speech community cannot partially overlap. In fact, if we allow for the possibility of an innovation partially diffusing through a proto-language (as discussed in the previous section), then this is what we would often expect.

3.2 Reading diachrony in glottometric diagrams

J&L criticize glottometric diagrams on the grounds that they are "pure data display" and thus carry no diachronic information (other than the trivial information that all the lects displayed are descended from the same proto-language). While it is true that glottometric diagrams do not directly show a temporal dimension, we believe that, like trees, they encapsulate hypotheses regarding the relative chronology of lineage-splitting events.

This becomes apparent once we realize that a glottometric diagram not only summarizes the innovations that have taken place in a language family, but also represents the mutual intelligibility relations among the dialects of the proto-language: the more isoglosses there are that connect a group of dialects, and the more "subgroupy" these isoglosses are, the more mutually intelligible those dialects were.

Let us define a "language" as a set of dialects connected by links of mutual intelligibility and disconnected from other dialects. In a dialect continuum, what are initially mutually intelligible lects diversify progressively from each other by

3. We admit that this point would be clearer if we referred to the innovations in our data as "*potential* shared innovations" whose status as "shared" or "parallel" would need to be determined by the degree to which they are supported by other potential shared innovations. We thank J&L for highlighting this potential source of confusion.

undergoing local innovations. For a while, the innovations taking place inside the continuum increase the difference across dialects yet do not jeopardize their overall mutual intelligibility. In spite of its emerging internal fragmentation, the language as a whole may remain alive for a period and still undergo its own innovations on a larger scale. In a glottometric diagram, this is shown by the coexistence of isoglosses of smaller scope (change affecting local dialects) with those of global scope (change affecting the language as a whole).

As changes accumulate over time, the links of mutual intelligibility among the dialects become weaker and ultimately disappear (with the weakest links disappearing first). In visual terms, this is equivalent to successively removing the weakest isoglosses from the glottometric diagram. As the weakest isoglosses are successively removed, the glottometric diagram "breaks apart" into disconnected sets of isoglosses; this is equivalent to the proto-language breaking up into mutually unintelligible daughter languages. This approach constitutes one possible way to formalize the notion of *linkage breaking* (Ross 1997: 222–228) – we can observe how a former dialect network progressively breaks into smaller dialect networks in a recursive manner until we reach the languages that are currently observable. This process is partly reminiscent of the diachrony that can be read in a tree diagram – as subgroups split successively into smaller subgroups – except that the initial stages of the evolution involve a dialect continuum showing intersecting isoglosses, a situation which is itself incompatible with a tree as traditionally conceived.

In more formal terms, we propose the following definitions:

i. A **glottometric diagram** is a *weighted hypergraph* (see J&L's footnote 14) whose nodes are dialects, whose edges are isoglosses (i.e., sets of dialects defined by one or more exclusively-shared innovations), and whose edge weights are the subgroupiness values of these isoglosses.
ii. A **language** is a *connected component* of such a weighted hypergraph – in other words, a set of dialects that are chained together by isoglosses and disconnected from any other dialects in the diagram.
iii. The **chronology** of a language family is found by successively removing the weakest edges from the hypergraph and at each stage noting how the dialects are partitioned into connected components (i.e., into languages).

We can illustrate these ideas using a glottometric diagram of the languages of North Vanuatu (Figure 1). The glottometric map in Figure 2 (from François 2017: 72) plots the same results onto a geographical map.[4]

4. For details of the dataset and methodology used to produce these diagrams, see François (2014), Kalyan & François (2018). From left to right in Figure 1 (Northwest to Southeast in Figure 2), the language names read as follows: HIW Hiw, LTG Lo Toga, LHI Lehali, LYP Löyöp, VLW Volow, MTP Mwotlap, LMG Lemerig, VRA Vera'a, VRS Vurës, MSN Mwesen, MTA Mota, NUM Nume, DRG Dorig, KRO Koro, OLR Olrat, LKN Lakon, MRL Mwerlap.

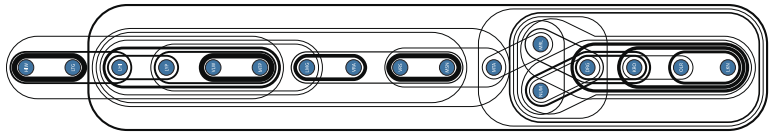

Figure 1. Glottometric diagram of the Oceanic languages of North Vanuatu

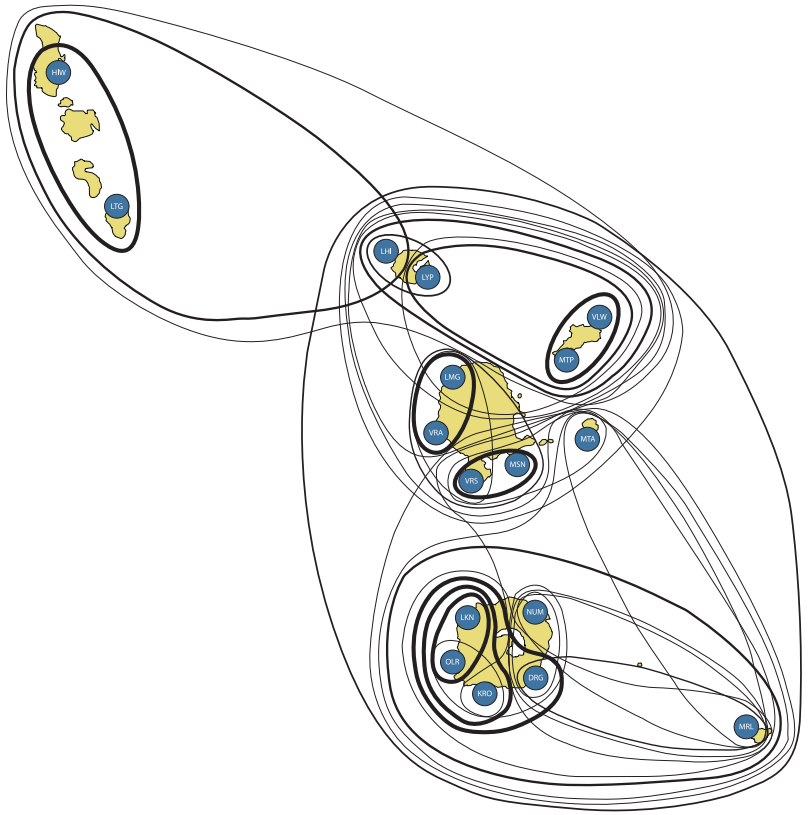

Figure 2. Glottometric map of the Oceanic languages of North Vanuatu

As can easily be seen, the isoglosses covering these 17 languages form a single connected set, reflecting the fact that the dialects that they descend from were all initially mutually intelligible. Due to the strength of the isogloss connecting the three westernmost dialects ({HIW-LTG-LHI}), all of the dialects remain connected even if the 21 weakest isoglosses are removed from the diagram. But as soon as we remove the 22nd-weakest isogloss (namely, {HIW-LTG-LHI}), the glottometric diagram breaks apart, as shown in Figure 3.

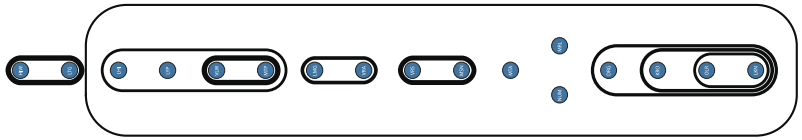

Figure 3. Glottometric diagram of the Oceanic languages of North Vanuatu, with the 22 weakest isoglosses removed

Here, the two westernmost dialects (the Torres group, {HIW-LTG}) have split off from the remainder (the Banks group). We interpret this as the first break in mutual intelligibility that occurred among the languages of North Vanuatu, resulting in a division of the original proto-language into two mutually-unintelligible intermediate proto-languages. Continuing in this manner, we drop the next weakest isogloss (i.e., the next-weakest link of mutual intelligibility), namely the one encompassing the Banks group. Figure 4 shows the situation that must have prevailed when that subgroup broke up, i.e., when the various dialects within it lost mutual intelligibility.

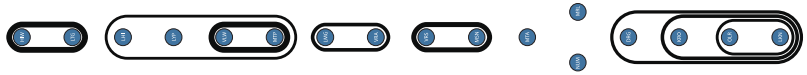

Figure 4. Glottometric diagram of the Oceanic languages of North Vanuatu, with the 23 weakest isoglosses removed

The result was then a set of eight languages, i.e. eight sets of dialects ({HIW-LTG}, {LHI-LYP-VLW-MTP}, {LMG-VRA}, {VRS-MSN}, {DRG-KRO-OLR-LKN}, {MTA}, {NUM}, and {MRL}), which, for some time, continued to evolve each as a single language community – as indicated by the amount of shared innovations characterizing each one. The process of *linkage breaking* illustrated here continued for several generations, eventually leading up to the different languages we know today.

In sum, Historical Glottometry does make it possible to infer a chronology of lineage-splitting events;[5] this becomes apparent once we realize that a glottometric diagram not only provides a synoptic overview of the innovations that occurred across the history of a language family, but also provides a map of the mutual-intelligibility relations among the dialects of the proto-language.

4. Conclusion

We hope to have shown that Historical Glottometry does not challenge the family tree model once *incomplete lineage sorting* has been taken into account. Our approach is actually meant as a critique of the tree model as traditionally understood, where an innovation must necessarily affect the *whole* of the population in which it occurs. The crucial observation of the wave model – that innovations frequently overlap – pertains to the developments that occur in a proto-language before it splits up (i.e., when it is still a network of mutually intelligible dialects). Thus, this observation is naturally captured within the framework of "incomplete lineage sorting" proposed by J&L; the only difference in our argument is that we allow for "variation" in the proto-language that is not only morphological or sociolinguistic, but also areal (dialectal). Finally, glottometric diagrams, far from being "pure data display," do in fact encode information about the order in which lineage splits most likely occurred.

This is not to say that we see no room for further improvement and refinement in Historical Glottometry. In particular, unlike standard approaches in computational phylogenetics, we do not currently have a generative model that could be used for estimating dates of innovations and population splits. One approach would be to apply existing models of incomplete lineage sorting (e.g. Pagel & Meade 2004; Wen, Yu & Nakhleh 2016) to linguistic data.[6] Another approach, which we are currently pursuing, is to directly develop a generative model of the spread of linguistic innovations in a network, adapting Madigan & York's (1995) work on Bayesian graphical models of discrete data. We believe that many promising avenues of research open up once we look beyond the restrictive assumptions of the family-tree model as traditionally understood – regardless of whether we draw our inspiration from the wave model of Schmidt (1872) and Schuchardt (1900) or from the insights of phylogenetic systematics (Maddison 1997; Galtier & Daubin 2008).

5. This type of inference should be used with caution, however: the inferred chronology is dependent on the ranking of isoglosses by subgroupiness, which itself may be sensitive to small changes in the data (particularly for low subgroupiness values).

6. See Verkerk (this issue) for an initial attempt in this direction.

Acknowledgements

This work contributes to the research program "Investissements d'Avenir," overseen by the French National Research Agency (ANR-10-LABX-0083): LabEx *Empirical Foundations of Linguistics*, Strand 3 – "Typology and dynamics of linguistic systems."

References

Baum, David A. & Stacey D. Smith. 2013. *Tree Thinking: An Introduction to Phylogenetic Biology*. New York: Macmillan.

Bloomfield, Leonard. 1933. *Language*. London: George Allen & Unwin.

Brugmann, Karl. 1884. Zur Frage nach den Verwandtschaftsverhältnissen der indogermanischen Sprachen. *Internationale Zeitschrift für allgemeine Sprachwissenschaft* 1, 226–256.

Donohue, Mark. 2013. Who Inherits What, When? Toward a Theory of Contact, Substrates, and Superimposition Zones. *Language Typology and Historical Contingency: In Honor of Johanna Nichols* ed. by Balthasar Bickel, Lenore Grenoble, David A. Peterson & Alan Timberlake, 219–240. (= *Typological Studies in Language*, 104.) Amsterdam: John Benjamins. https://doi.org/10.1075/tsl.104.09don

François, Alexandre. 2011. Social Ecology and Language History in the Northern Vanuatu Linkage: A Tale of Divergence and Convergence. *Journal of Historical Linguistics* 1:2.175–246. https://doi.org/10.1075/jhl.1.2.03fra

François, Alexandre. 2014. Trees, Waves and Linkages: Models of Language Diversification. *The Routledge Handbook of Historical Linguistics* ed. by Claire Bowern & Bethwyn Evans, 161–189. London: Routledge.

François, Alexandre. 2017. Méthode comparative et chaînages linguistiques: Pour un modèle diffusionniste en généalogie des langues. *Diffusion: implantation, affinités, convergence* ed. by Jean-Léo Léonard, 43–82. (= *Mémoires de la Société de Linguistique de Paris*, XXIV.) Louvain: Peeters.

Galtier, Nicolas & Vincent Daubin. 2008. Dealing with Incongruence in Phylogenomic Analyses. *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 363:1512.4023–4029. https://doi.org/10.1098/rstb.2008.0144

Greenhill, Simon J. & Russell D. Gray. 2012. Basic Vocabulary and Bayesian Phylolinguistics: Issues of Understanding and Representation. *Diachronica* 29:4.523–537. https://doi.org/10.1075/dia.29.4.05gre

Kalyan, Siva & Alexandre François. 2018. Freeing the Comparative Method from the Tree Model: A Framework for Historical Glottometry. *Let's Talk About Trees: Genetic Relationships of Languages and Their Phylogenic Representation* ed. by Ritsuko Kikusawa & Lawrence A. Reid, 59–90. (= *Senri Ethnological Studies*, 98). Ōsaka: National Museum of Ethnology.

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

Maddison, Wayne P. 1997. Gene Trees in Species Trees. *Systematic Biology* 45:523–536. https://doi.org/10.1093/sysbio/46.3.523

Madigan, David & Jeremy York. 1995. Bayesian Graphical Models for Discrete Data. *International Statistical Review* 63:2.215–232.

Nichols, Johanna. 1992. *Linguistic Diversity in Space and Time*. Chicago: University of Chicago Press. https://doi.org/10.7208/chicago/9780226580593.001.0001

Pagel, Mark & Andrew Meade. 2004. A Phylogenetic Mixture Model for Detecting Pattern-Heterogeneity in Gene Sequence or Character-State Data. *Systematic Biology* 53:571–581. https://doi.org/10.1080/10635150490468675

Ross, Malcolm D. 1997. Social Networks and Kinds of Speech-Community Event. *Archaeology and Language 1: Theoretical and Methodological Orientations* ed. by Roger Blench & Matthew Spriggs, 209–261. London: Routledge.

de Saussure, Ferdinand. 1995 [1916]. *Cours de linguistique générale*. Paris: Éditions Payot & Rivages.

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

Schuchardt, Hugo. 1900. *Über die Klassifikation der romanischen Mundarten. Probe-Vorlesung, gehalten zu Leipzig am 30. April 1870*. Graz.

Wen, Dingqiao, Yun Yu & Luay Nakhleh. 2016. Bayesian Inference of Reticulate Phylogenies Under the Multispecies Network Coalescent. *PLOS Genetics* 12:5.e1006006. https://doi.org/10.1371/journal.pgen.1006006

Address for correspondence

Siva Kalyan
Department of Linguistics
School of Culture, History and Language
College of Asia and the Pacific
Australian National University
9 Fellows Road
Acton, ACT 2601
Australia
siva.kalyan@anu.edu.au

Co-author information

Alexandre François
CNRS-ENS-LaTTiCe
Australian National University
alexandre.francois@ens.fr