Méthode comparative et chaînages linguistiques
Pour un modèle diffusionniste en généalogie des langues

Alexandre FRANÇOIS
CNRS-LACITO — Australian National University

Abstract

Linguistic diffusion is commonly equated with contact, and contrasted with genealogy. This article takes a new perspective, by showing how diffusion lies in fact at the heart of language genealogy itself. Indeed, the Comparative method has taught us to identify genetic subgroups based on sets of shared innovations; but each of these innovations necessarily had to diffuse from speaker to speaker across a network of then mutually intelligible idiolects. Such a diffusionist approach to language genealogy allows us to model language change as it really took place in the social and geographical space of past societies.

Crucially, the entangled isoglosses typical of dialect continuums and linkages (Ross 1988) cannot be handled by the Tree model, which is solely based on divergence; but they are easily captured by a diffusionist approach such as the Wave model, where the key process is convergence. After comparing the theoretical underpinnings of these two models, I introduce Historical Glottometry, a new quantitative approach aiming to free the Comparative Method from any cladistic assumption, and to reconcile it with a wave-based analysis. Finally, data from a group of Oceanic languages from Vanuatu illustrate the powerful potential of Glottometry as a new method for linguistic subgrouping.

1 DIFFUSION ET GÉNÉALOGIE

1.1 La diffusion, moteur de la généalogie des langues

En linguistique historique, il est d'usage d'associer la notion de *diffusion* aux phénomènes aréaux, tels qu'ils se manifestent à travers les processus d'emprunt ou de convergence structurale entre langues distinctes, y compris entre langues non apparentées[1]. La diffusion linguistique, conçue alors comme synonyme de *contact*, sera ainsi contrastée avec la transmission intergénérationnelle des langues, laquelle opère au sein d'une même communauté linguistique. Il est même devenu fréquent d'opposer les deux dimensions comme si elles étaient orthogonales : c'est ainsi qu'on décrira les relations

[1] Je remercie mes collègues Siva Kalyan, Malcolm Ross et Matthieu Segui pour leurs suggestions au cours de la rédaction de cet article. Ces travaux ont d'abord été présentés, notamment, à la Conférence Internationale de Linguistique Historique (ICHL21 à Oslo, ICHL22 à Naples). Ils s'insèrent dans le programme « Investissements d'Avenir » géré par l'Agence Nationale de la Recherche ANR-10-LABX-0083 (Labex EFL), plus précisément dans son axe *Typology and dynamics of linguistic systems*.

de parenté au sein d'une famille linguistique comme des processus de « transmission verticale », alors que les phénomènes de contact seront décrits comme reflétant la « transmission horizontale » (Cavalli-Sforza & Feldman 1981 ; Wang & Minett 2005 ; Currie *et al*. 2010).

Ce type de contraste est particulièrement présent dans les travaux comparatistes visant à reconstruire la parenté des langues. Au moment d'appliquer la méthode comparative pour identifier des groupes généalogiques[2], on prend soin de séparer ce qui relève de la parenté linguistique *stricto sensu*, et ce qui relève de phénomènes aréaux (cf. Aikhenvald & Dixon 2001). Les points communs entre deux langues apparentées seront jugés pertinents s'ils relèvent de l'*évolution interne* des systèmes, et de structures transmises par héritage dit « vertical » ; mais ils seront exclus de l'analyse si l'on peut montrer qu'ils sont dûs à des processus ultérieurs de contact, dits de « diffusion » : ces derniers processus ne sont pas censés entrer en ligne de compte dans la structure généalogique d'une famille. Ainsi, Labov (2007:347) définit la diffusion linguistique comme le « transfert [de propriétés linguistiques] d'une branche à l'autre dans un arbre généalogique » (« transfer across branches of the family tree »).

Pourtant, de telles formulations omettent un point crucial : que *la généalogie des langues reflète toujours elle-même un processus de diffusion linguistique*.

Quand bien même on déciderait de s'en tenir strictement aux relations de parenté en excluant le contact, on n'en aurait pas fini avec la diffusion. En effet, la méthode comparative définit chaque unité généalogique sur la base d'une liste d'*innovations exclusivement partagées* entre ses membres. Or ces innovations ne se sont pas produites d'un seul coup : quelles qu'elles soient, elles ont forcément dû se propager de locuteur à locuteur avant de se stabiliser au sein d'une communauté linguistique. On a bel et bien affaire à de la diffusion. Pour la distinguer de la diffusion entre langues distinctes (dite « contact »), j'emploierai ici le terme de DIFFUSION INTERNE : interne à une langue, c-à-d. à un réseau d'idiolectes mutuellement intelligibles.

La thèse principale du présent article est que l'on ne peut analyser la parenté au sein d'une famille que si on la traduit en termes de diffusion d'innovations au sein d'un réseau social. En d'autres termes, je propose une approche diffusionniste de la généalogie des langues.

1.2 Deux modèles contradictoires : l'Arbre et les Ondes

Il ne s'agit pas simplement d'un ajustement terminologique, mais de revoir en profondeur notre manière de concevoir l'histoire des familles linguistiques. Ce changement de perspective sera ici argumenté à travers la comparaison de deux modèles concurrents pour représenter la généalogie des langues : d'un côté, le modèle arborescent ou cladistique (*Stammbaum*) ; de l'autre, le modèle dialectologique, également connu sous le nom de théorie des ondes (*Wellentheorie*).

Depuis Schleicher (1853) et les néogrammairiens, la vision la plus répandue de la généalogie linguistique est celle d'une longue série de scissions, à mesure qu'une

[2] Au terme *génétique*, employé souvent en linguistique historique, est ici substitué *généalogique* (cf. Haspelmath 2004:222). Ainsi, ce que l'anglais appellerait *genetic subgroup* sera ici désigné *groupe* (ou groupement) *généalogique*. Pour une définition de la généalogie linguistique, voir §3.3.

langue-mère se divisait en plusieurs langues-filles. Le modèle de l'arbre est entièrement bâti autour de cette idée que la diversification des langues repose principalement sur un processus de divergence, lequel sera volontiers corrélé à des événements de scission de population (Campbell 2004:165).

Or cette vision simpliste de l'histoire des langues présente bien des faiblesses, dont certaines ont été reconnues très tôt. C'est ainsi que Schmidt, puis Schuchardt, ont inauguré dès les années 1870 une longue tradition de critiques théoriques et méthodologiques de ce modèle arborescent. Parmi les défauts que l'on doit reconnaître au modèle cladistique, je m'arrêterai surtout sur l'un d'entre eux : son incapacité à concevoir la possibilité de *groupes généalogiques entrecroisés*. Pourtant, on peut aisément montrer [cf. §3.3] que l'évolution interne des familles de langues donne spontanément lieu à des groupes entrecroisés, en sorte qu'une langue C peut appartenir à une unité génétique ABC en même temps qu'à une unité BCD ou CDE. C'est le phénomène bien connu des continuums dialectaux et de ce que j'appellerai *chaînages linguistiques*, devant lesquels le modèle de l'arbre se révèle impuissant.

Face aux impasses du modèle cladistique, la présente étude propose de réhabiliter le modèle alternatif dit de la « théorie des ondes », précisément conçu par Schmidt et Schuchardt. Considéré à une époque comme le modèle de référence (cf. Saussure 1916, Bloomfield 1933:317), ce modèle a fini par être négligé en linguistique historique, à partir du milieu du XX^{e} s. – excepté en dialectologie. À l'inverse de l'arbre généalogique, le modèle des ondes met l'accent sur les phénomènes de diffusion linguistique d'un locuteur à l'autre (convergence) – et non pas sur la divergence, qui n'est qu'un épiphénomène. Et contrairement aux idées reçues, ce contre-modèle est parfaitement compatible avec la rigueur de la méthode comparative, indispensable au travail de reconstruction historique.

La section 2 présentera d'abord le modèle de l'arbre tel qu'il est typiquement utilisé en linguistique historique, et en examinera les présupposés ainsi que les limites. La section 3 montrera que la généalogie des langues est avant tout un processus de diffusion, et expliquera les conséquences importantes de cette idée. Enfin, la section 4 proposera de formaliser et modéliser cette approche diffusionniste de la généalogie des langues à l'aide d'une méthode quantitative : la Glottométrie historique.

2 LE MODÈLE CLADISTIQUE

2.1 Les principes du modèle arborescent

Examinons d'abord la manière dont les arbres sont classiquement construits, et interprétés. Soit cinq langues modernes, nommées K, L, M, N, O. Ces langues sont dites apparentées si elles remplissent certains critères : en particulier, un nombre non négligeable de morphèmes et lexèmes clairement apparentés, pas nécessairement ressemblants entre eux, mais présentant des correspondances phonétiques régulières en nombre suffisant pour que ces liens ne puissent s'expliquer ni par le hasard, ni par des processus d'emprunts.

Dire que les parlers K, L, M, N, O sont apparentés implique qu'ils descendent d'un ancêtre commun – une « proto-langue », que l'on appellera ici proto-KLMNO. Ceci peut être représenté sous la forme de la *Figure 1*, structure en « rateau » ou *polytomie* : chaque langue y apparaît comme un descendant autonome de la proto-langue, sans que ne soit proposée aucune structure interne pour la famille.

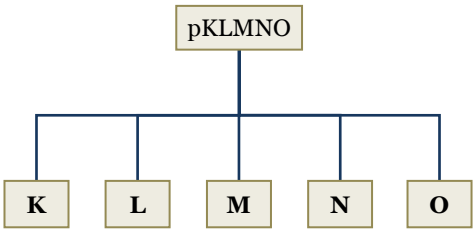


Figure 1 – Un arbre généalogique non structuré

Ce type d'arbre plat correspond parfois à une situation historique réelle, comme dans le cas où une société ancestrale s'est rapidement dispersée en sous-communautés distinctes, avec rupture rapide du contact. Ainsi, lorsque la civilisation antique dite Lapita s'est dispersée, vers 3100 BP, à travers les archipels du Pacifique, sa langue le proto-océanien a éclaté en plusieurs sous-familles non ordonnées entre elles (Pawley 1999 ; cf. §4.3.1). Dans d'autres cas, un diagramme comme la *Figure 1* est simplement dû à la difficulté de reconstruire la structure interne d'une famille, du fait de données insuffisantes ou ambiguës. Le type de structure que les diachroniciens espèrent reconstituer est plutôt celle qui identifie les affinités internes au sein de la famille, et regroupe entre elles les langues qui partagent un ancêtre plus récent. Ce type d'arbre idéal est illustré dans la *Figure 2*.

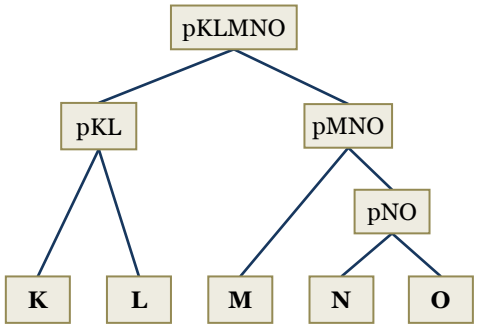


Figure 2 – Un arbre généalogique structuré

Un tel arbre synthétise un ensemble d'hypothèses concernant la structure interne de la famille. Ainsi, on y affirme que les langues K et L forment ensemble un groupe généalogique KL, par contraste avec M, N et O qui forment un autre groupe MNO ; au sein de ce dernier, on affirme que N et O forment leur propre groupe distinct de M. Les groupes sont enchâssés les uns dans les autres : N appartient au groupe NO, qui à son tour forme une « branche » dans le groupe MNO.

Ces hypothèses pourraient être formulées en termes purement classificatoires sans référence au temps. Cependant, la pratique usuelle est d'interpréter ces diagrammes cladistiques comme un scénario historique : on dira alors que la succession des nœuds dans l'arbre, de haut en bas, suit l'ordre chronologique des événements historiques. Une autre idée répandue – si simpliste soit-elle (Pulgram 1961) – est que chaque nœud dans l'arbre réfère à une communauté linguistique discrète, en sorte qu'une fourche reflèterait la division d'un groupe social en deux communautés séparées.

Dans ce contexte, dire que les trois langues M, N, O forment ensemble un groupe à part au sein de leur famille (un *clade* < gr. κλάδος 'branche') signifierait qu'elles

descendent toutes d'une seule protolangue intermédiaire, dite proto-MNO. Suite à la scission de la communauté ancestrale proto-KLMNO, la langue proto-MNO se serait développée séparément de proto-KL. À l'appui d'une telle reconstruction, on citera un ensemble d'innovations linguistiques de divers types (phonologiques, grammaticales, lexicales, etc.) qui se trouveraient reflétées dans les langues modernes M, N, O, et uniquement elles.

En effet, si M, N et O partagent une certaine propriété linguistique, cela peut être dû (1) à un héritage de leur lointain ancêtre commun (proto-KLMNO) ; (2) à une innovation parallèle ; (3) à un emprunt récent entre elles ; ou enfin, (4) à une innovation linguistique qui aurait eu lieu une seule fois dans l'histoire, à l'époque où les ancêtres de M, N et O constituaient (des variétés mutuellement intelligibles d') une seule et même langue. C'est ce dernier cas, et lui seul, qui justifie de reconstituer un état de langue intermédiaire commun à ces trois langues M N O, qu'on appellera proto-MNO. Reconstruire le système du proto-MNO implique alors d'identifier les innovations partagées exclusivement par les trois membres M N O, à l'exclusion des autres membres de la famille[3]. Ce principe des *innovations exclusivement partagées* (IEP) constitue, depuis Leskien (1876:VII), la pierre angulaire de tout travail visant à identifier des groupes généalogiques au sein d'une famille.

Plus le nombre de ces IEP est grand, plus solide sera notre certitude que la langue proto-MNO correspond bien à une réalité historique, et donc que la branche ⟨MNO⟩ de cette famille est légitime. Pour prendre un nouvel exemple d'Océanie, la trentaine de langues polynésiennes modernes présentent entre elles d'innombrables points communs, reflétant des innovations que l'on ne retrouve pas dans les 450 autres langues de la grande famille océanienne. Ce grand nombre d'IEP donne une grande certitude à l'existence, dans l'histoire de la famille océanienne, d'un groupe généalogique polynésien nettement distinct des autres. Si l'on traduit cette conclusion au plan des populations, on dira que la linguistique historique rend très plausible l'existence d'une communauté « proto-polynésienne » qui se serait développée plus ou moins à l'écart des autres langues océaniennes – quelque part entre Tonga et Samoa – pendant une assez longue période (cf. Pawley & Ross 1995).

Le raisonnement est récursif. La proto-langue pMNO sera définie par une série d'innovations partagées ; mais la *Figure 2* montre également que N et O forment aussi un sous-groupe, lequel sera défini par un nouvel ensemble d'innovations partagées entre ces deux langues, à l'exclusion de M. Le modèle de l'arbre présume que les innovations propres au proto-NO doivent être chronologiquement postérieures à celles du proto-MNO : c'est aussi ce que montre la structure verticale de l'arbre, avec ses nœuds en cascade. Nous le verrons, cet axiome – selon lequel les groupes larges doivent forcément avoir connu leurs innovations *avant* les groupes en leur sein – est l'un des aprioris problématiques du modèle cladistique [§3.4, 4.3.5].

2.2 L'arbre, un modèle fondé exclusivement sur la séparation

Dans l'interprétation classique des arbres, chaque nœud est censé correspondre à une communauté particulière, qui se serait développée à l'écard des autres (Fox 1995:123).

[3] C'est l'équivalent de ce que la phylogénétique cladistique appelle les *synapomorphies* (Lecointre & Le Guyader 2001:16, Page & Holmes 2009).

La séparation en question est souvent comprise comme un événement réel de division sociale tel qu'une migration, processus par lequel une société autrefois unifiée se scinde en deux groupes séparés, avec perte de contact. D'autres cas sont aussi possibles – comme l'isolement dû à l'intrusion d'autres langues ; ou le fractionnement progressif d'anciens réseaux de communication, par lequel une communauté se fragmenterait en plusieurs groupes distincts.

Afin d'obtenir une structure arborescente robuste telle qu'illustrée dans la *Figure 2*, dotée de nœuds intermédiaires (par contraste avec la structure plate de la *Figure 1*), le processus de scission sociale doit s'être réitéré à plusieurs reprises au cours des siècles. Chaque événement de séparation doit avoir été suivi d'une période de stabilité – ne serait-ce que quelques générations – pour permettre aux innovations linguistiques d'émerger, se propager et se stabiliser au sein de la nouvelle communauté (Pawley & Ross 1995), avant que ne survienne une nouvelle séparation.

Cette importance donnée à la scission et à la divergence est centrale au modèle arborescent ; c'est à la fois sa qualité et son défaut. C'est une qualité, au sens où le modèle de l'arbre peut nous aider à reconstruire des événements de division sociale lorsqu'ils ont vraiment eu lieu, et peut les représenter à l'aide d'un diagramme visuellement très simple. Mais c'est également le principal défaut de l'approche cladistique, car elle nous force à reconstruire des événements de scission même lorsqu'ils ne se sont en réalité jamais produits. C'est ce qu'on appelle un *artefact du modèle* : du fait des aprioris inhérents à sa structure, le modèle cladistique déforme la réalité historique en imposant un scénario unique (la scission) pour tous les processus de glossogénèse. Ce faisant, il rend indétectables les autres scénarios possibles – tels ceux qui impliquent une interaction continue entre dialectes, sans scission ni rupture de contact.

Tout bien réfléchi, il n'est au monde aucune population dont on puisse réduire l'histoire à une simple succession de scissions définitives – pourtant le seul scénario autorisé par le modèle de l'arbre. Certes, il existe des familles linguistiques qui ont connu de tels événements de séparation au cours de leur développement, sous la forme de migrations ou autres catastrophes de ce type ; mais ces divisions, corrélées avec des processus de divergence linguistique, sont toujours précédées ou suivies d'autres modes d'interaction sociale, dont les conséquences linguistiques – nous le verrons bientôt – ne sont pas compatibles avec une représentation arborescente.

2.3 La diffusion externe : le grand absent du modèle cladistique

Un point crucial pour notre discussion est le fait que, dans le modèle cladistique, une langue ne peut être située que sous un seul nœud de même niveau. Par exemple, si M appartient à une branche MNO, il ne peut pas appartenir en même temps à une branche KLM : l'un des axiomes du modèle arborescent est que les groupes de même niveau s'excluent mutuellement, et ne sont pas autorisés à se chevaucher. Cette idée semble assez logique si l'on interprète les nœuds dans un arbre comme des événements de scission sociale définitive : car si les populations parlant pKL et pMNO étaient en effet séparées avec perte de contact, alors on aurait du mal à concevoir comment une partie (M) des descendants de pMNO, et pas les autres (N, O), pourraient partager des innovations avec pKL.

Imaginons donc que, dans la *Figure 2*, un trait linguistique soit partagé par les langues L et M, et seulement ces deux langues. Comment interpréter une telle observation ? Dans une approche strictement cladistique, ceci constituerait un problème que

l'on ne peut résoudre qu'au prix d'hypothèses *ad hoc*, destinées à sauver la structure arborescente. Par exemple, on proposera que le trait commun soit en réalité un cas de *rétention partagée* (nommée *symplésiomorphie* en cladistique) à partir de l'ancêtre commun pKLMNO, propriété qui aura été perdue par les autres descendants (K, N, O) : dans ce cas, le trait commun ne serait diagnostique d'aucun lien généalogique entre L et M, si ce n'est leur appartenance générale au groupe KLMNO. Ou alors, on proposera que le trait est en effet une innovation, mais qu'il a émergé de manière indépendante en L et M, par innovation parallèle (*homoplasie*).

Enfin, une troisième hypothèse pourrait être que le trait a été innové de manière « interne » à une seule langue, par exemple L ; puis qu'il a été emprunté par une autre langue M par CONTACT entre L et M, après qu'elles se seraient constituées comme deux langues distinctes. Tout le monde admet que le contact entre langues constitue une cause puissante de changement linguistique ; cependant, dans une lecture stricte du modèle cladistique, ces phénomènes de contact ne sont pas censés avoir leur place dans un arbre généalogique (cf. la citation de Labov donnée en §1.1). Selon ce raisonnement, un trait de la langue L emprunté par M après leur séparation ne saurait être considéré comme un argument pour poser un groupe généalogique LM ; il sera décrit comme un effet du « contact », et jugé non-informatif du point de vue phylogénétique.

Plusieurs auteurs ont fait part de leur frustration vis-à-vis du modèle cladistique, du fait que les arbres ne savent représenter que la divergence linguistique, à l'exclusion des phénomènes de convergence par contact (cf. Fox 1995:124; Dixon 1997; Aikhenvald & Dixon 2001; Bossong 2009; Drinka 2013). Ces auteurs soulignent que les emprunts lexicaux ou grammaticaux, ou tout autre fait de diffusion, appartiennent à l'histoire des langues tout autant que le matériel hérité d'un ancêtre. À cette objection, les cladistes répondent que les arbres ne cherchent qu'à représenter une partie de l'histoire des langues, à savoir leur généalogie *stricto sensu*, et rien d'autre. Les autres faits d'évolution linguistique – notamment les effets du contact – doivent être traités par d'autres modèles (Labov 2007 ; Campbell & Poser 2008:327). Au passage, c'est là un point qu'il ne faut pas perdre de vue : la parenté des langues ne constitue qu'une portion de leur histoire, et l'on doit se garder d'accorder aux arbres plus d'importance qu'ils n'ont.

2.4 La diffusion interne : le talon d'Achille du modèle cladistique

Dans les pages qui suivent, l'argument que je développerai contre le modèle arborescent reprend partiellement l'objection que je viens de mentionner, mais avec une nuance importante. Ma thèse est que les arbres non seulement omettent de représenter le *contact* entre langues, mais également – et c'est bien plus gênant – qu'ils sont incapables de représenter correctement leurs relations de parenté. Cet argument s'appuie également sur les phénomènes de diffusion horizontale ; sauf qu'au lieu de diffusion externe (contact entre langues distinctes), il s'agit cette fois de prendre en compte les phénomènes de DIFFUSION INTERNE – autrement dit, la diffusion d'innovations entre les idiolectes mutuellement compréhensibles qui constituent une communauté linguistique.

Or le raisonnement qui précède à propos d'un trait linguistique qui serait partagé par les langues L et M nécessitera d'être totalement repensé s'il s'avère que les ensembles KL et MNO n'avaient en réalité jamais cessé de coexister. Certes, on peut concevoir que KL et MNO constituent deux ensembles dialectaux déjà partiellement différenciés ; mais s'ils demeurent mutuellement intelligibles et continuent d'interagir l'un avec l'autre, alors rien ne devrait empêcher de nouvelles innovations de se diffuser

d'un groupe à l'autre, par exemple de L à M, de part et d'autre de leur « frontière ». S'agit-il alors de contact ou de généalogie ? Si L et M sont encore des dialectes mutuellement intelligibles à ce moment-là, alors ils appartiennent à une seule et même « langue », en sorte qu'on a bien affaire ici à de la diffusion interne. Ces innovations partagées par L et M sont destinées à se transmettre à leurs descendants, exactement au même titre que celles qui s'étaient auparavant propagées entre K et L, ou entre M N O. Par conséquent, ces innovations entrecroisées définissent bel et bien la généalogie interne du groupe, sans qu'il soit légitime de les négliger sous prétexte qu'elles ne refléteraient que du « contact », un « transfert d'une branche à l'autre ».

Au bout du compte, on obtient ici trois groupes généalogiques au sens de la méthode comparative, définis chacun par leurs innovations exclusivement partagées : un groupe K-L, un groupe L-M, un groupe M-N-O. On hésite à parler de clades ou de branches (branche KL, branche LM, branche MNO), car on entrevoit déjà la difficulté d'une représentation cladistique. Pourtant cette configuration, dans laquelle les groupements généalogiques s'entrecroisent, n'est pas si difficile à concevoir à l'esprit : il s'agit tout simplement d'une structure de type *chaîne dialectale*, elle-même un cas particulier de continuum, ou de ce que nous allons appeler *chaînage de langues* [§3.5]. Or ce type de configuration généalogique pourtant simple est incompatible avec le modèle réducteur de l'arbre. Véritable talon d'Achille de la phylogénétique cladistique, les situations de continuums et de chaînages nécessitent de recourir à un modèle non arborescent.

3 LES ONDES DE DIFFUSION À LA SOURCE DE LA GÉNÉALOGIE

Loin de s'opposer l'une à l'autre, la diffusion et la transmission intergénérationnelle constituent deux aspects indissociables de la généalogie des langues. Les progrès récents de la sociolinguistique au cours des dernières décennies nous permettent de repenser les principes mêmes de la phylogénétique des langues.

3.1 La fiction de la langue et les réseaux d'idiolectes

Les néo-grammairiens assignaient le changement linguistique à une entité abstraite qu'ils appelaient la « langue » (ou « proto-langue »). La conception même du modèle arborescent repose sur le préjugé que la *langue* formerait une unité atomique allant de soi, un nœud distinct des autres nœuds. Ainsi, si l'on observe que trois langues modernes M N O reflètent toutes trois une même innovation, on jugera plus économique – plus « parcimonieux » – de considérer que ce changement se sera produit *une seule fois*, dans *une seule langue* (le proto-MNO) plutôt que séparément dans trois parlers différents. Un corollaire de ce raisonnement est la tentation de réifier les ensembles dialectaux caractérisés par des traits communs, comme formant une seule et même « langue » à part des autres membres de sa famille.

Cette fiction fut mise à mal dès la fin du XIX^{e} s. par les premiers travaux de dialectologie (Gilliéron 1880, Wenker 1881), lesquels montrèrent que toute langue est en réalité un réseau plus ou moins homogène de dialectes. Il devint vite évident que les traits linguistiques sont distribués dans l'espace de manière parfois complexe, représentée visuellement à l'aide d'*isoglosses*. Loin de s'aligner toujours parfaitement, ces isoglosses définissent généralement des régions différentes du réseau social, et se chevauchent (cf. Saussure 1985 [1916]:261-289, Chambers & Trudgill 1998:91).

Ces apports de la dialectologie ont ensuite été confirmés par les études de sociolinguistique. Celles-ci ont montré que l'innovation linguistique n'a pas lieu d'un seul

coup dans une « langue », mais qu'elle prend toujours la forme de changements individuels, lesquels se propagent d'un locuteur à l'autre par imitation au cours de leurs échanges quotidiens (cf. Labov 1963, 1994, 2001, 2007 ; Milroy & Milroy 1985 ; Milroy 1987). Ces travaux ont aussi montré que la distribution de ces traits innovants devait non seulement être définie dans l'espace géographique (*diatopique*) mais également dans l'espace social (*diastratique*), comme marqueurs d'appartenance à des classes ou des groupes spécifiques de la société.

Si l'on veut comprendre la notion de changement linguistique, la seule unité d'observation qui soit valide n'est donc pas la « langue » ni même le « dialecte », mais l'IDIOLECTE – c-à-d. la compétence linguistique d'un locuteur individuel à un moment donné. Les dialectes et les langues ne sont au fond rien d'autre que des *réseaux d'idiolectes mutuellement intelligibles* – réseaux plus ou moins homogènes. Lorsque les diachroniciens parlent d'une innovation *x*>*y* qui aurait eu lieu « une seule fois » dans « une langue », ils résument en une formule simpliste ce qui est en réalité un long processus de diffusion le long de vastes réseaux d'idiolectes, parfois sur plusieurs générations.

3.2 Innovation et propagation

La diffusion du changement linguistique à travers les réseaux d'idiolectes repose sur un processus d'imitation de locuteur à locuteur, parfois appelé *accommodation* (Street & Giles 1982; Trudgill 1986; Giles & Ogay 2007). Certains auteurs parlent de *propagation* (Croft 2000), d'autres d'*épidémiologie linguistique* (Enfield 2003, 2008).

Une innovation linguistique commence toujours par émerger dans le discours de certains individus, sous la forme d'une nouvelle manière de s'exprimer – qu'il s'agisse de changements phonétiques, lexicaux, phraséologiques… Pour peu que mon auditeur soit séduit par cette innovation et y voie un moyen possible d'assurer le succès de ses objectifs de communication, il pourra l'adopter – inconsciemment le plus souvent – dans sa propre pratique discursive. Ce faisant, sans s'en rendre compte, il contribuera ainsi à relayer et propager cette innovation dans son propre réseau social.

Durant un temps plus ou moins long, la prononciation ou tournure innovante connaîtra une phase de compétition avec la norme précédente, laquelle offrira plus ou moins de résistance (cf. François 2011a:204-210). C'est ainsi, par exemple, que le français *second* est menacé depuis des siècles par *deuxième* ; que *Pas de problème* subit la concurrence, depuis une quinzaine d'années, de *Pas de souci* ; que les occlusives dentales devant voyelle antérieure sont de plus en plus souvent palatalisées (ex. [vwaˈtyʁ] > [vwaˈʧyʁ] 'voiture')… Si elle remporte ce bras-de-fer, l'innovation linguistique deviendra statistiquement dominante, et s'établira comme la nouvelle norme de tout un groupe social (propriété d'une langue, d'un dialecte ou d'un sociolecte). Par la suite, ce trait linguistique innovant sera transmis « verticalement » aux générations suivantes de locuteurs, au même titre que le reste du système hérité.

C'est cette diffusion des innovations au sein d'un réseau d'idiolectes (au sein d'une même « langue ») qui sous-tend les relations généalogiques auxquelles s'intéressent les études phylogénétiques. Sachant qu'un groupe généalogique est défini par un ensemble d'innovations linguistiques, il faut d'abord que ces innovations se soient propagées de locuteur à locuteur au sein d'un ancien réseau social, avant de pouvoir se transmettre par héritage. Diffusion horizontale et transmission verticale sont deux composantes également essentielles de la généalogie des langues.

3.3 Quand les groupes généalogiques s'entrecroisent

À mesure qu'il s'établit dans une portion du réseau idiolectal, chaque changement linguistique définit sa propre isoglosse – généralement une zone géographiquement contiguë, représentable sur une carte, indiquant l'aire où l'innovation se sera propagée et fixée.

Dans un continuum linguistique caractérisé par l'intelligibilité mutuelle entre dialectes adjacents, la situation normale est que les isoglosses se chevauchent constamment. Ainsi, soit huit dialectes proches nommés A—H, et six innovations numérotées #1 à #6 (*Figure 3*). Imaginons que l'innovation #1 soit née dans le dialecte D avant de se propager aux dialectes adjacents jusqu'à couvrir la zone *CDE*; que #2 ait ensuite couvert *AB*; que #3 ait affecté *CDEF*; #4 *FG*; #5 *EF*, et #6 *EFGH*.

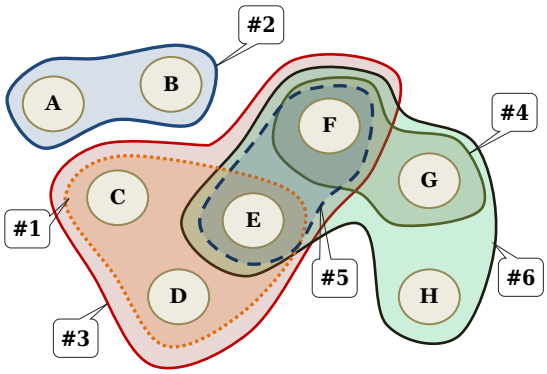


Figure 3 – Isoglosses entrecroisées dans un continuum dialectal ou chaînage de langues

Les premières innovations ayant ciblé les dialectes C-D-E n'auront pas altéré ces variétés assez radicalement pour rompre leur intelligibilité avec d'autres dialectes voisins. Ainsi, en l'absence de barrière physique ou de séparation migratoire totale, rien n'empêche la prochaine innovation de cibler un groupe E-F, puis une autre F-G, et ainsi de suite. Chaque innovation constitue un cas de CONVERGENCE (pour les dialectes qui participent ensemble à cette innovation, par ex. F et G en #4) tout autant qu'elle forme un cas de DIVERGENCE (pour les dialectes qui se retrouvent différenciés à la faveur de ce changement, comme G et H en #4) – cf. François (2011a:231).

Les diverses innovations accumulées au fil du temps laissent leur empreinte dans chaque dialecte local. Considérons deux dialectes, par exemple F et G. D'un côté, les changements qu'ils partagent l'un avec l'autre (#4, #6) auront accru leur similarité sur certains aspects de leurs systèmes ; de l'autre côté, les changements qui n'auront affecté que l'un d'entre eux (soit seul, soit lui et d'autres dialectes voisins – par ex. #3, #5) auront accentué leur différence. À force d'accumuler des innovations divergentes (c-à-d. non partagées), les deux dialectes F et G, initialement intelligibles, finiront par devenir opaques l'un à l'autre au point de constituer deux langues distinctes.

À moins d'un processus ultérieur de nivellement dialectal de type *koinè*, chaque membre du continuum héritera dans son patrimoine génétique l'ensemble des innovations auxquelles ses ancêtres auront participé, et les transmettra à ses descendants. En ce sens, toutes les innovations dont il est ici question, illustrées dans la *Figure 3*, définissent la structure généalogique de la famille, au sens de la méthode comparative. Définissons ainsi la notion de *groupe généalogique* :

(1) On appelle GROUPE GÉNÉALOGIQUE un ensemble de langues apparentées dont les ancêtres ont autrefois partagé une ou plusieurs innovations linguistiques par diffusion, à une époque où ces variétés étaient mutuellement intelligibles.

Rien dans cette définition ne suggère que les groupes généalogiques doivent être discrets ou imbriqués les uns dans les autres, sans intersection. Il est en fait parfaitement commun, dans le monde réel, que les groupes généalogiques se chevauchent, comme dans la *Figure 3*. En d'autres termes, on a tort de chercher à représenter les structures généalogiques des langues à l'aide d'une structure arborescente.

3.4 La théorie des ondes

C'est un raisonnement théorique similaire qui se trouve aux fondements de la théorie des ondes, ou *Wellentheorie*, développée par Hugo Schuchardt et Johannes Schmidt dans les années 1870 (Schmidt 1872, Schuchardt 1900 [1870]). Il s'agissait pour eux de contrer le modèle arborescent (*Stammbaumtheorie*) de leur aîné August Schleicher.

En réalité, ces auteurs allaient plus loin encore, puisqu'ils rejetaient en bloc la méthode comparée : Schuchardt, par exemple, remettait en cause la foi des néogrammairiens en la régularité des changements phonétiques (Schuchardt 1885). Cette radicalité a porté tort au modèle des ondes, car elle l'a associé à un rejet pur et simple de la méthode comparative, par ailleurs si puissante et si convaincante. À mon avis, une synthèse est possible entre ces deux approches : je propose de préserver la force de la méthode comparative – fondée notamment sur le principe de régularité – tout en remplaçant le fallacieux modèle de l'arbre par une approche plus réaliste et empirique, inspirée par la théorie des ondes.

Au fondement de cette théorie, se trouve l'observation que tout changement linguistique commence d'abord par apparaître à un point, puis se propage de locuteur en locuteur à travers la communauté linguistique. Cette propagation est comparable à une « onde » qui part d'un centre et se déplace vers l'extérieur (cf. Bloomfield 1933: 317). Chaque onde, correspondant à la diffusion d'une innovation, est en principe indépendante des autres ondes, qu'elle peut parfaitement chevaucher. Telle était d'ailleurs la vision de Ferdinand de Saussure (1985 [1916]:289] :

> « Puisque les aires d'innovation varient d'étendue d'un cas à l'autre, deux idiomes voisins peuvent avoir une particularité commune sans former un groupe à part dans l'ensemble, et chacun d'eux peut être relié aux idiomes contigus par d'autres caractères. »

C'est exactement ce que montre la *Figure 3* ci-dessus : les idiomes E et F ont en commun certaines particularités sans pour autant former un groupe à part ; et chacun est relié aux autres idiomes par d'autres caractères. En d'autres termes, l'existence d'un groupe généalogique EF défini par certaines innovations n'empêche pas l'existence de groupes entrecroisés CDE ou FG.

Autre corollaire de cette approche diffusionniste : une innovation affectant un petit groupe de dialectes (ex. #4 en FG) peut être suivie dans le temps par une autre innovation affectant un plus grand ensemble de dialectes (ex. #6 en EFGH)[4]. À propos

[4] Pour des exemples empiriques de tels processus, voir par exemple Garrett (2006) pour les dialectes grecs, Geraghty (1983) pour les dialectes fidjiens, François (2011a:201) pour le Vanuatu. Voir aussi notre discussion finale en §4.3.5.

de l'espace roman, Chambon (2011:299) appelle ce type de configuration « changements en accordéon »,

> « faisant se succéder alternativement des innovations caractérisant des espaces très réduits (…) avec des innovations suprarégionales, (…) se diffusant sur de vastes espaces »

Là encore, une telle configuration serait incompatible avec le modèle de l'arbre, qui force une chronologie allant toujours des groupes larges (EFGH) vers les groupes étroits (FG) [§2.1]. La théorie des ondes n'impose pas de tels aprioris, et demeure ouverte à bien d'autres scénarios.

3.5 Des continuums dialectaux aux chaînages de langues

Le problème de l'intersection des isoglosses a toujours été central à la dialectologie (cf. Bloomfield 1933:321). Il n'est donc pas surprenant que les dialectologues, qui cherchent à observer avec précision la distribution des traits linguistiques dans l'espace, tendent à préférer le modèle des ondes – ou un modèle dérivé de ce dernier – plutôt que les schémas arborescents. Les réseaux de dialectes romans, flamands, arabes ou chinois – pour ne prendre que quelques exemples – ne pourront jamais être adéquatement rendus par un arbre.

On pourrait proposer que les deux modèles sont en fait « complémentaires » : les arbres seraient idoines pour représenter les relations généalogiques entre LANGUES distinctes ; tandis que les ondes seraient un outil adapté aux relations complexes entre DIALECTES, au sein de chaque langue. On dirait alors que les deux modèles sont utiles, mais à différents grains d'observation. Malgré son apparence de bon sens, une telle suggestion oublie un point essentiel : c'est que la plupart des familles de langues dans le monde sont justement issues d'anciens continuums dialectaux. Dans la mesure où les anciennes innovations locales sont transmises fidèlement aux générations qui suivent, les langues qui en résultent devraient garder trace de toutes ces isoglosses enchevêtrées. Or, si les arbres échouent à représenter correctement les relations généalogiques entre dialectes, ils échoueront nécessairement tout autant à représenter les liens de parenté entre les langues qui en sont issues.

Ce dernier argument a été notamment avancé par Malcolm Ross, et a donné lieu au concept de *linkage* (Ross 1988, 1996, 1997, 2001) – notion que je propose de traduire par *chaînage de langues*. Ross (1988:8) définit un chaînage comme un « groupe de lectes apparu par différenciation dialectale ». Chaque fois qu'un continuum dialectal, structuré comme dans la *Figure 3* ci-dessus, voit diminuer l'intelligibilité entre ses membres, il devient un chaînage. Un chaînage est donc une famille de langues dont les membres sont liés entre eux par une série d'innovations enchevêtrées ; c'est une famille dont la généalogie interne est incompatible avec une représentation cladistique.

Ross avait initialement développé cette notion dans le contexte des langues d'Océanie occidentale (Western Oceanic) dont il reconstruisait l'histoire linguistique ; cependant, on peut aisément la généraliser au reste du monde. Bien des familles ou sous-familles dans le monde sont manifestement des chaînages linguistiques – et ce, qu'elles aient été décrites ou non à l'aide de ce terme. Les langues océaniennes de Fidji (Geraghty 1983) ou de Polynésie (Gray, Bryant & Greenhill 2010), le groupe karnique du Pama-Nyungan (Bowern 2006), les langues athapascanes du nord (Krauss & Golla 1973, Holton 2011) ; certaines portions des familles sémitique (Huehnergard & Rubin 2011) ou sinitique (Hashimoto 1992, Chappell 2001) ; les langues indo-aryennes de la région Kamta en Inde (Toulmin 2009), les langues indo-iraniennes (Korn 2003), les

langues romanes (Penny 2000:9–74; Ernst *et al.* 2003-2008) ou germaniques (Ramat 1998) – voire la famille indo-européenne dans son ensemble (Bloomfield 1933:316; Anttila 1985:305; Garrett 2006; Drinka 2013)… : ce sont là divers exemples, parmi bien d'autres dans le monde, de familles linguistiques dont on a pu montrer qu'elles ont connu une accumulation d'innovations au cours du temps, présentant tant d'intersections qu'elles sont incompatibles avec le modèle arborescent.

3.6 L'arbre, un cas particulier de chaînage

À en juger par les études empiriques sur les familles du monde, il faut se rendre à l'évidence : les familles dotées d'une généalogie authentiquement arborescente sont bien plus rares dans le monde réel que les tenants du modèle cladistique ne veulent l'admettre. On pourrait même se demander, en toute bonne foi, si ce modèle de l'arbre possède une quelconque légitimité pour représenter la généalogie des langues.

D'aucuns pourraient proposer de sauver ce dernier au moins pour les quelques (sous-) familles avec lesquelles il serait, soi-disant, compatible. Une telle idée serait conforme à l'attitude de juste milieu qu'on lit parfois, selon laquelle le modèle de l'arbre et celui des ondes sont décrits comme « complémentaires », et tous deux d'égale légitimité (Hock 1991:454; Rankin 2003:186; Labov 2007). Une fois de plus, cette conclusion ne semble pas justifiée.

Alors que le modèle cladistique est incapable de représenter correctement une structure en chaînage, l'inverse n'est pas vrai : l'approche par ondes ou isoglosses permet de détecter et de représenter une structure dite arborescente (« tree-like ») chaque fois que c'est approprié. En effet, une telle structure ne serait, en réalité, qu'un cas particulier de *chaînage* – à savoir, un cas exceptionnel dans lequel les isoglosses seraient par hasard dépourvues d'intersection, et ordonnées chronologiquement de la plus large à la plus étroite.

Imaginons ainsi que, dans la *Figure 3* ci-dessus, les membres du groupe AB ne partagent aucune innovation avec les autres membres de la famille : c'est ce que traduit l'absence de toute isoglosse englobant soit A, soit B, soit AB avec d'autres langues. Une telle observation peut s'expliquer, par exemple, par une situation d'isolement géographique ou social, par exemple suite à une migration. Dans ce cas, la configuration qu'on obtient est justement le type de scission que les arbres cherchent en permanence à représenter.

En somme, l'existence occasionnelle de telles scissions dans les sociétés du monde n'est pas une raison suffisante pour justifier qu'on préserve le modèle cladistique. Pour revenir à notre exemple, il est vrai que l'existence d'un groupe AB séparé du reste pourrait être représentée visuellement, si l'on veut, par une « branche » qui relierait l'ancêtre commun (proto-ABCDEFGH) à proto-AB ; mais qu'en serait-il du reste de la famille ? L'enchevêtrement des isoglosses observées dans le groupe CDEFGH demeurerait incompatible avec un arbre, et requerrait de toute façon une approche non-arborescente. Tout bien réfléchi, un diagramme isoglossique comme la *Figure 3* est à la fois nécessaire et suffisant pour représenter visuellement les scissions lorsqu'elles existent, et un arbre n'apporterait rien de plus.

4 LA GLOTTOMÉTRIE HISTORIQUE : MODÉLISER LES CHAÎNAGES

Il nous faut donc définir une méthode susceptible d'identifier et de représenter les groupes généalogiques même lorsqu'ils s'entrecroisent. Après un aperçu de quelques

modèles non cladistiques (§4.1), la section finale de cette étude présentera en détail une méthode possible pour formaliser la théorie des ondes : la Glottométrie historique.

4.1 Quelques contre-modèles pour la généalogie des langues

Si les arbres sont restés si populaires en linguistique historique en dépit des vices de conception qu'on leur a très vite reconnus, c'est notamment pour une raison triviale : ils proposent une représentation visuellement élégante et facile à lire d'un scénario très simple. Pour qu'un jour soit réhabilitée une approche plus réaliste fondée sur la dialectologie, il importe donc de concevoir un modèle qui se présente de manière aussi lisible et intuitive, sans pour autant y sacrifier l'exactitude des faits. Or depuis un siècle, plusieurs contre-modèles ont été proposés face au *Stammbaum* de Schleicher.

Au cours des dernières années, diverses études phylogénétiques se sont posé le problème des groupes génétiques à faible signal : les auteurs y ont mis à contribution des méthodes bayésiennes fondées sur le maximum de vraisemblance (« Bayesian maximum-likelihood methods ») pour tenter d'évaluer le degré de solidité de chaque nœud dans un arbre (cf. Dunn *et al.* 2008; Greenhill & Gray 2009; Greenhill, Drummond & Gray 2010; Gray, Bryant & Greenhill 2010; Bowern & Atkinson 2012 ; Dunn 2014). L'intérêt de ces méthodes est d'éviter une lecture simpliste des arbres généalogiques, et de proposer des outils empiriques – quoique généralement limités au lexique – pour jauger la validité des hypothèses cladistiques dans chaque famille. L'ennui, c'est que ces approches phylogénétiques ne réussissent pas à s'affranchir du modèle cladistique. Confrontées à une famille de type *chaînage*, elles chercheront à quantifier son degré d'*arborescence* (« tree-likeness »), et pourront éventuellement calculer à quel point cette famille est éloignée d'une structure cladistique. Cependant, ces méthodes continuent de présenter leurs résultats sous la forme de cladogrammes – tantôt un seul arbre, tantôt un grand nombre d'arbres potentiels pour une même famille (Dunn 2014: 201-204). Elles ne cherchent jamais à imaginer une alternative radicale à l'approche cladistique elle-même, qu'elles ne remettent d'ailleurs jamais en question. Au bout du compte, elles ne permettent pas de modéliser la véritable structure généalogique des chaînages de langues, car elles ne se donnent pas les moyens de penser et de représenter les chevauchements de groupes.[5]

Plus prometteuses sont les représentations en réseaux proposées par Ross (1997: 223, 234) ou Forster *et al.* (1998:185) ; ou encore les NeighborNets, qui sont de plus en plus reconnus comme une approche alternative aux arbres (Bryant *et al.* 2005; Heggarty *et al.* 2010). Ces derniers sont capables de représenter des distances deux-à-deux (*pairwise distance*) entre taxons, sous la forme de groupements entrecroisés. Les réseaux NeighborNets permettent ainsi de saisir graphiquement les enchevêtrements complexes que constituent parfois les familles de langues dans la réalité[6].

Enfin, parmi les autres tentatives pour modéliser la diversité linguistique, il faut citer la *dialectométrie* (Séguy 1973; Guarisma & Möhlig 1986; Goebl 2006; Nerbonne 2010;

[5] Un autre problème est que certains de ces travaux de phylogénétique ne s'appuient pas sur la méthode comparative. Ainsi, Dunn *et al.* (2008) identifie des groupements généalogiques non pas à partir d'innovations partagées, mais à partir d'une simple liste de propriétés typologiques telles que l'ordre des mots.

[6] Pour une présentation critique des réseaux NeighborNet comme représentation de la généalogie des langues, voir François (2014:179).

Szmrecsanyi 2011). Cet ensemble de méthodes cherche à visualiser les distances deux-à-deux entre membres d'un continuum dialectal, à l'aide de calculs portant sur de vastes bases de données. Les résultats prennent typiquement la forme de cartes choroplèthes colorées, utilisant les distances chromatiques pour représenter les distances linguistiques. Cette approche dialectométrique, si inspirante soit-elle, ne vise cependant pas à représenter les relations généalogiques au sens de la méthode comparative : conformément aux pratiques usuelles en dialectologie, les distances linguistiques n'y sont généralement évaluées qu'à l'aune des traits synchroniques, sans distinguer les rétentions des innovations partagées.

4.2 Une synthèse entre la méthode comparative et la théorie des ondes : La Glottométrie historique

Le reste de cette étude vise à présenter une synthèse des principes théoriques élaborés dans les pages précédentes, et à définir un nouveau modèle baptisé *GLOTTOMÉTRIE HISTORIQUE* (cf. François 2014, Kalyan & François *ss presse*). Cette méthode entend conjuguer la précision et le réalisme des approches dialectologiques – en particulier la dialectométrie, qui lui a inspiré son nom – avec la puissance de raisonnement de la méthode comparative.

L'objectif de la Glottométrie historique est d'identifier des groupements généalogiques dans une famille de langues, et d'en calculer la solidité relative en observant avec précision la répartition des isoglosses dans l'espace. La glottométrie part du principe que des groupes généalogiques entrecroisés peuvent parfaitement coexister au sein d'un continuum, mais que certains de ces groupes seront plus saillants que d'autres. Alors que le modèle de l'arbre impose une lecture en tout-ou-rien (tel groupe généalogique existe ou n'existe pas), la glottométrie conçoit des groupes à solidités relatives. Par exemple, dans la *Figure 3* ci-dessus, imaginons que les langues E-F aient partagé 60 innovations, et F-G seulement 12 : une telle mesure linguistique permettrait de montrer, de manière empirique et quantitative, que le groupe généalogique EF est plus « solide » que le groupe FG ; non pas pour dire que ce dernier n'existe pas, mais simplement qu'il est pourvu d'un degré d'attestation historique moindre. Des liens linguistiques solides entre deux communautés pourront alors être interprétés comme le signe de relations sociales intenses, révélant par exemple l'étroitesse des liens entre les communautés E et F, plutôt qu'entre F et G. C'est bien là le type de résultat que souhaitent obtenir les historiens des langues et des populations.

Le modèle qu'il nous faut définir doit capturer l'accumulation des événements historiques qui ont donné lieu à l'émergence de nouvelles langues à partir d'un ancêtre commun. Par conséquent, nous devons porter attention non pas aux propriétés synchroniques des langues, mais à celles qui reflètent des *innovations* : il s'agit de respecter le principe énoncé par Brugmann (1884:231), l'un des piliers de la méthode comparative. On notera que le rejet du modèle arborescent de Schleicher n'empêche aucunement, par ailleurs, de suivre rigoureusement les principes de la méthode comparative en ce qui concerne l'identification des innovations. C'est ainsi, par exemple, qu'il faut interpréter la *Figure 3* ci-dessus : chaque isoglosse[7] y correspond à une ou plusieurs *innovations partagées* – lesquelles ont besoin de la méthode comparative pour être identifiées.

[7] Les dialectologues emploient le terme « isoglosse » indépendamment de son statut historique, comme rétention ou innovation. Précisons que les isoglosses dont il est question en Glotto-

…/…

Les procédures permettant d'identifier les innovations sont celles de la méthode comparative – et sont donc les mêmes que celles que l'on emploie dans l'approche cladistique. Ces procédures reposent notamment sur le principe de *régularité des changements phonétiques* (*pace* Schuchardt 1885), sur la notion de directionalité de certains changements, et sur les principes de chronologie relative, par exemple. La méthode comparative permet également de distinguer entre changements phonétiques réguliers et irréguliers ; ceci peut fournir de précieux indices pour identifier, d'une part, les changements précoces au sein d'un continuum dialectal, et d'autre part, les changements acquis ultérieurement par contact entre langues bien après leur séparation (cf. Biggs 1965 pour le rotumien). De tels outils sont précieux pour identifier les données de la *diffusion interne* [§2.4] qui sous-tendent les relations généalogiques entre langues.

Après avoir identifié un certain nombre d'innovations, on est en mesure d'énumérer les langues qu'elles ont affectées. Chaque groupe présente un nombre ε, celui de ses « innovations exclusivement partagées » : ce taux fournit une mesure de la fréquence avec laquelle ses membres tendaient à s'imiter les uns les autres ; on obtient ainsi une première estimation de la force de leurs liens sociaux.

La Glottométrie historique propose de mesurer la solidité relative des groupes généalogiques dans des situations de chaînage à l'aide d'outils plus sophistiqués encore – en particulier, des calculs de *cohésion* et de *solidité*. Ces notions sont présentées dans la section suivante, à travers une étude de cas portant sur des langues océaniennes du Vanuatu.

4.3 Étude glottométrique d'un chaînage du Vanuatu

4.3.1 Le chaînage du Vanuatu septentrional

Le Vanuatu, archipel de Mélanésie insulaire situé au sud du Pacifique, compte un total de 138 langues distinctes (François *et al.* 2015). Rapporté à une population de 0,23 millions d'habitants, ce nombre donne au Vanuatu la plus forte densité linguistique au monde (ibid.). Toutes ces langues descendent du proto-océanien (POc) – langue austronésienne, ancêtre commun d'environ 500 langues océaniennes dans le Pacifique insulaire, parlée par les premiers hommes qui colonisèrent le Vanuatu il y a environ 3100 ans (Pawley 1999).

Si l'on met de côté trois langues polynésiennes qui sont arrivées dans l'archipel au cours des derniers siècles, les 135 autres langues forment ensemble un vaste chaînage linguistique (Tryon 1996, Lynch 2000:181, François 2011b, François *et al.* 2015:11) : leur diversité moderne reflète trois millénaires de fragmentation dialectale *in situ*, sans apport extérieur identifiable. Ce que l'on peut reconstituer initalement comme un vaste réseau linguistiquement homogène a par la suite connu, au fil des millénaires, une accumulation d'innovations locales, prenant la forme de multiples isoglosses enchevêtrées les unes dans les autres : c'est ainsi que les dialectes du proto-océanien se sont diversifiés au point de former la mosaïque de langues distinctes que nous connaissons aujourd'hui.

Parmi toutes les langues du Vanuatu, dix-sept sont parlées dans les îles Torres et Banks au nord de l'archipel, par une population qui a toujours connu des traditions de

métrie historique sont toujours des ISOGLOSSES HISTORIQUES – comme dans Bloomfield (1933: 316) ou Anttila (1989:305).

(2) ‘voler’: HIW *βeneɣ*; LTG *βəneɣ*; LHI *pɔl*; LYP *pɪl*; VLW ^{m}*bɛl*; MTP ^{m}*bɛl*; LMG *pœl*; VRA ^{m}*bɔl*; VRS ^{m}*bœl*; MSN *pɔl*; MTA *pal*; NUM ^{m}*bal*; DRG ^{m}*baːl*; KRO ^{m}*bɛ͡al*; OLR *pal*; LKN *pal*; MRL ^{m}*bɑl*.

La connaissance de l’histoire phonologique de la région permet de déterminer que les formes des îles Torres (HIW *βeneɣ*; LTG *βəneɣ*) reflètent régulièrement **panako* ‘voler’, l’étymon reconstruit au niveau du proto-océanien (cf. Blust 2013). S’il est vrai que ces deux formes ont connu des changements phonétiques, elles sont lexicalement conservatrices : elles constituent un cas de rétention partagée (*symplésiomorphie*), et ne donnent donc aucune information de nature généalogique.

À l’inverse, les quinze formes des îles Banks reflètent toutes un étymon que l’on peut reconstruire, sur la base des correspondances phonétiques régulières, sous la forme *^{m}*balu* (François 2005:493). Sachant que cet étymon est inconnu ailleurs dans la grande famille océanienne, il constitue donc une innovation lexicale locale, commune aux quinze langues des îles Banks. Au passage, une telle affirmation n’oblige aucunement à réifier une langue « proto-Banks » unitaire, qui soit séparée du reste de sa famille comme le serait un nœud dans un arbre : on a simplement affaire ici à un ensemble de quinze dialectes, lesquels auront partagé entre eux certaines innovations, à l’époque où ils étaient encore mutuellement intelligibles[9].

L’identification d’une innovation implique que les états de caractère puissent être chronologiquement ordonnés. Dans ce cas précis, il est aisé de déterminer que **panako* précède *^{m}*balu* dans le temps : c’est donc ce dernier qui constitue l’innovation. Cette procédure implique parfois de raisonner sur la chronologie relative des changements, chaque fois que les données le justifient [§4.3.5.1].

Au moment d’identifier les innovations, pour éviter tout arbitraire, il est bon de résister à la tentation de juger si l’on a affaire à une innovation « banale » ou « rare ». Certes, ce type de précaution est rendue nécessaire dans une approche en tout-ou-rien comme le modèle arborescent : lorsqu’on construit un arbre, une innovation qualitativement *rare* (ex. un changement phonétique, syntaxique ou sémantique inhabituel ou paradoxal) peut fournir un contre-exemple fatal à une hypothèse de regroupement généalogique – alors qu’un contre-exemple *banal* (un changement assez fréquent typologiquement) sera balayé du revers de la main, comme étant probablement dû à une innovation parallèle [cf. §2.3]. Mais ce type de raisonnement, toujours spéculatif et *ad hoc*, devient superflu dans un modèle capable de gérer les isoglosses entrecroisées. L’objection peut même aller plus loin. Imaginons qu’un groupe AB repose sur dix innovations qualitativement rares et BC sur dix innovations banales ; il ne me semble même pas justifié de considérer qu’AB soit un groupe plus *solide* que BC : ces deux groupes devraient être mis à égalité, indépendamment du statut – banal ou rare – des innovations qui les définissent. On dira simplement que la communauté B a partagé autant d’innovations linguistiques avec A qu’avec C.

Dans le même esprit, on évitera de chercher à décider si telle innovation partagée entre deux langues est véritablement le résultat d’une diffusion inter-dialectale, ou s’il pourrait s’agir éventuellement d’une innovation parallèle (*homoplasie*). Ce type de raisonnement demeure lui aussi toujours spéculatif, et peu fiable. Mon hypothèse, qui

[9] D’ailleurs, cet ensemble-là est entrecoupé par certaines isoglosses, comme nous le verrons dans le *Tableau 1* ci-dessous. Voir la discussion finale en §4.3.5.

s'est avérée payante, est de miser plutôt sur une quantité importante de données en incluant toutes les isoglosses que l'on peut reconstruire, sans chercher à les trier. Si les données possèdent un signal généalogique fort, celui-ci émergera de lui-même dans les calculs, et fera apparaître comme négligeables les éventuels cas d'innovations parallèles.

4.3.3 UNE BASE DE DONNÉES COMPARATIVE

En somme, la clef pour obtenir des résultats de qualité est d'abord de mettre en place une vaste base de données d'innovations historiques ancrées dans la géographie linguistique. S'agissant des langues du nord du Vanuatu, j'ai ainsi compilé une base de 474 différentes innovations. Ces dernières incluent 21 cas de changement phonétique régulier (localisé au niveau du système phonologique) ; 116 cas de changement phonétique irrégulier (localisé au niveau de chaque lexème individuel) ; 91 cas de changement morphologique ; 10 de changement syntaxique, et 236 de remplacement lexical.

Pour chaque langue L et chaque innovation *i*, la base indique '1' lorsque L reflète *i*; '0' lorsque l'on peut positivement affirmer[10] que L n'a pas subi *i* ; et un blanc '–' quand les données ne permettaient pas de trancher dans un sens ou dans l'autre. Au total, la base contient 8058 points : 2728 positifs ('1'), 5040 négatifs ('0') et 290 agnostiques ('-'). Le *Tableau 1* fournit un échantillon de dix innovations extraites de la base, et en montre la distribution parmi les dix-sept membres du chaînage. Chaque innovation est ici identifiée à l'aide d'un nombre (1e colonne) et d'un mot-clef (2e colonne).

id		HIW	LTG	LHI	LYP	MTP	VLW	LMG	VRA	VRS	MSN	MTA	NUM	DRG	KRO	OLR	LKN	MRL
1	*^{m}*balu*	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
2	**late*	1	1	1	0	0	0	1	1	1	0	0	0	0	0	0	0	0
3	**suRi*	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0
4	*$o^{ŋ}$*ga*	0	0	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
5	**ira(ŋ)*	0	0	0	1	1	1	1	1	0	0	0	0	0	0	0	0	0
6	*t>ʔ	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0
7	**one*	0	0	0	0	0	0	1	1	1	1	–	1	1	1	0	1	0
8	**wo*	0	0	0	0	0	0	0	0	1	1	1	1	1	1	0	0	0
9	**ŋoRo*	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1
10	**tolira*	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0

Tableau 1 – Quelques isoglosses dans les langues des Banks et Torres (échantillon)

[10] Un '0' implique parfois que la langue est conservatrice sur ce point, ayant préservé l'état initial de la proto-langue ; mais il peut également signifier que la langue considérée ne reflète simplement pas l'innovation *i* – sachant qu'elle peut parfaitement en refléter une autre. Ainsi, dans la ligne 5 du *Tableau 1*, les langues présentant un '0' sont celles qui ne reflètent pas le changement **ura(ŋ)* → ****i****ra(ŋ)*. Parmi elles, certaines ont préservé **ura(ŋ)*, et d'autres reflètent des innovations distinctes – ex. KRO, OLR, LKN reflètent un changement **ura(ŋ)* → **uraŋ****i***, non montré ici (François 2011a:198).

Les dix innovations du *Tableau 1* sont définies comme suit :

1. [*^{m}*balu*] REMPLACEMENT LEXICAL
   POc **panako* 'voler' a été remplacé par un nouveau verbe *^{m}*balu* [§4.3.2]
2. [**late*] CHANGEMENT PHONÉTIQUE PROPRE À UN LEXÈME
   **late* 'scinder' a modifé sa voyelle en **lete*
   Ex. VRA *lɪʔ* est un reflet régulier de **lete*, et non de **late*.
3. [**suRi*] MORPHOLOGIE
   Le verbe POc **suRi* 'suivre' (cf. Lichtenberk 2013) a été grammaticalisé en préposition dative. Ex. MTP *hij*, reflet régulier de **suRi*, marque le datif (François 2001:683).
4. [*$o^{ŋ}$*ga*] CHANGEMENT PHONÉTIQUE PROPRE À UN LEXÈME
   POc *$wa^{ŋ}$*ga* 'bateau' a changé irrégulièrement en *$o^{ŋ}$*ga*
   Ex. LMG *ɔk* est un reflet régulier de *$o^{ŋ}$*ga*, et non de *$wa^{ŋ}$*ga*.
5. [**ira*] CHANGEMENT PHONÉTIQUE PROPRE À UN LEXÈME
   **ura(ŋ)* 'écrevisse' (<POc **quraŋ*) a changé irrégulièrement en **ira(ŋ)* (François 2011a: 200). Ex. LYP *n-i͡ɛj* est un reflet régulier de **ira(ŋ)*, et non de **ura(ŋ)*.
6. [**t>ʔ*] CHANGEMENT PHONÉTIQUE SYSTÉMATIQUE
   **t* s'est changé régulièrement en occlusive glottale **ʔ*
7. [**one*] CHANGEMENT PHONÉTIQUE PROPRE À UN LEXÈME
   **eno* 'être allongé' (<POc **qenop*) s'est métathésé en **one*
   Ex. LMG *œn* est un reflet régulier de **one*, et non de **eno*.

   Note: L'étymon **qenop* a entièrement disparu en langue mota, où 'être allongé' est une forme non apparentée *rsa*. Ce remplacement lexical empêche d'évaluer empiriquement si le pré-mota avait préservé la forme originelle **eno* (→ codage '0') ou bien subi la métathèse en **one* comme les langues voisines (→ codage '1'). Dans un tel cas, je choisis de rester agnostique, et encode ce point comme un cas où la présence de l'innovation ne peut être ni prouvée ni infirmée (→ codage '-'). La Glottométrie historique attribue un statut particulier pour ces cas, qu'elle traite comme différents à la fois de 0 et de 1.
8. [**wo*] MORPHOLOGIE
   le clitique innovant **wo* a remplacé l'article nominal **na* pour les noms aliénables non-humains (François 2007)
9. [**ŋoRo*] REMPLACEMENT LEXICAL
   POc **matiruR* 'dormir' a été remplacé par **ŋoRo*, étymologiquement 'ronfler'.
10. [**tolira*] MORPHOLOGIE
    métathèse du pronom triel de 3^{e} personne : *ira-tolu → *tolira
    Ex. LKN *tɪlɪː* est un reflet régulier de **tolira* et non de **iratolu* (François s/p)

On notera que toutes les innovations considérées ici ont peu de chances de représenter des emprunts récents, et correspondent très probablement à des processus anciens de diffusion interne, datant de l'époque où les ancêtres de ces langues étaient des lectes mutuellement intelligibles : dans ce sens, elles permettent bien de reconstituer la généalogie interne de cette famille, au sens strict de la méthode comparative. C'est le cas, par exemple, des cas de remplacement lexical que j'ai mentionnés : d'abord, parce qu'ils portent sur du vocabulaire dit « fondamental », moins susceptible d'emprunt (Haspelmath & Tadmor 2009:65-68) ; ensuite, parce que le changement lexical a manifestement précédé le changement phonétique qu'ont connu ensuite ces mêmes langues (par ex., le lakon *ŋɔː* 'dormir' reflète **ŋoRo*, avec amuïssement régulier de **R* et allongement compensatoire, cf. François 2011b:150).

Les cas de *changement phonétique propre à un lexème* sont également un indice fort de la généalogie, car il est rare que ce type de changement se diffuse d'une langue à l'autre : ces altérations arbitraires de la forme d'un lexème ont tendance à ne se propager qu'entre des individus qui s'identifient comme locuteurs de la même langue au moment du changement (Ross 1988:12; François 2011a:200). Ce type de changement – appelé parfois *changement phonétique irrégulier* au sens où il n'est pas localisé dans le système phonologique, mais dans un lexème spécifique – a donc un fort pouvoir diagnostique, et mérite qu'on y prête attention dans les recherches généalogiques[11].

Comme le montre le tableau, la distribution des isoglosses historiques parmi les langues des Banks et Torres présente en permanence des intersections, à la manière de la *Figure 3* présentée en §3.3 : on a bel et bien affaire à un chaînage, qui ne saurait être représenté adéquatement par un arbre.

4.3.4 L'ANALYSE GLOTTOMÉTRIQUE

La définition de *groupe généalogique* donnée en §3.3 comporte un corollaire : c'est que n'importe quelle isoglosse historique peut, potentiellement, correspondre à une unité généalogique. Le problème est alors le risque d'obtenir pléthore de tels groupes, dont certains ne seraient que faiblement attestés, voire chimériques – dus par exemple à des cas de convergence accidentelle : c'est ce que l'on appellerait du « bruit » dans les résultats, alors que nous cherchons à détecter un « signal » généalogique.

Aussi importe-t-il de pouvoir jauger la solidité de ces groupes, en fonction de leurs attestations. C'est justement le but de l'analyse glottométrique, que de faire émerger les distributions d'isoglosses les plus significatives, afin de dégager l'histoire généalogique de chaque famille de langues.

4.3.4.1 La cohésion des groupes généalogiques

Pour tout groupe de langues, on peut calculer ce qu'on appellera sa ***cohésion***, une estimation de la solidarité entre ses membres.

Pour tout groupe G, appelons ***p*** le nombre d'isoglosses partagées par tous ses membres (exclusivement ou non), et qui confirment donc leur solidarité linguistique. À l'inverse, appelons ***q*** le nombre d'isoglosses qui « coupent » ce groupe G (c-à-d. qui impliquent une partie des membres de G, avec au moins un non-membre) : celles-ci contredisent la solidarité linguistique interne à G. On calcule pour ce groupe de langues un *taux de cohésion k* (en angl. *cohesiveness*) :

$$k_G = \frac{nombre\ d'innovations\ en\ commun}{nombre\ total\ d'innovations\ pertinentes} = \frac{\boldsymbol{p}}{(\boldsymbol{p}+\boldsymbol{q})}$$

Considérons, par exemple, le groupe formé par les deux langues lemerig et vera'a. Sur les 474 innovations listées dans la base, elles en partagent 134 (→ lignes 1, 2, 5, 6, 7 dans le *Tableau 1*) – y compris 9 qu'elles partagent exclusivement (→ ligne 6). À eux seuls, ces chiffres sont peu informatifs sur les relations réciproques entre ces deux

[11] De tels changements phonétiques sont plus fréquents qu'on ne le dit, y compris dans les langues dont le lecteur sera plus familier. Par exemple, le latin **laxāre* 'relâcher' est reflété régulièrement en français (*laisser*) ou en italien (*lasciare*) ; mais tout un ensemble de parlers romans du sud-ouest présentent une consonne initiale non-étymologique **d*, reflétant une proto-forme innovative **daxāre* : port./cat. ***d****eixar*, cast. ***d****ejar*, langued. ***d****aishar*, prov. ***d****eissar*…

langues : ce qui importe pour évaluer la cohésion interne du groupe, ce n'est pas un nombre absolu, mais le rapport entre le nombre d'innovations qui confirment sa solidarité (dans cet exemple, p=134) et celles qui la contredisent. Or, le lemerig partage 30 innovations avec des langues autres que le vera'a (ex. ligne 4 dans le *Tableau 1*); tandis que 14 sont partagées par le vera'a avec d'autres langues que le lemerig. La loyauté linguistique au sein de ce couple lemerig–vera'a s'est donc trouvée contredite q=44 fois. On calcule alors le taux de cohésion k du groupe :

$$k_{LMG-VRA} = 134/(134 + 44) \ = 0{,}75$$

En d'autres termes, sur les 178 innovations qui ont pu affecter le couple lemerig–vera'a ensemble ou séparément, 134 étaient partagées par les deux langues, soit 75%. Le lemerig et le vera'a présentaient donc une assez forte loyauté réciproque, chaque fois qu'il s'agissait d'adopter ou non une innovation.

On comparera ce taux à celui du groupe vera'a–vurës, sur la même île. Cette paire de langues reflète ensemble p=118 innovations, mais se trouve traversée par un total de q=88 isoglosses. On a donc :

$$k_{VRA-VRS} = 118/(118 + 88) = 0{,}57$$

De cette comparaison, on peut conclure que les ancêtres de la communauté vera'a entretenaient des liens linguistiques, et donc sociaux, plus forts avec la communauté linguistique lemerig au nord ($k_{LMG-VRA} = 75\%$), qu'avec les Vurës au sud ($k_{VRA-VRS} = 57\%$) – et ce, en dépit des liens sociaux forts existant aujourd'hui entre les sociétés de Vera'a et de Vurës. Le taux de cohésion fournit donc une métrique précieuse des relations généalogiques entre langues : cet outil permet d'inférer l'intensité des liens sociaux entre communautés anciennes, à partir des traces que leurs lectes ont laissées dans les langues modernes.

La base de données des innovations, structurée comme le *Tableau 1* [§4.3.3], peut ainsi servir de base pour calculer – grâce à un algorithme conçu par Siva Kalyan – les taux de cohésion pour l'ensemble des groupes attestés dans la région.

La plus forte cohésion observée est celle qui unit le volow et le mwotlap, avec 92% : sur toutes les innovations qui ont affecté l'une de ces deux langues, 92% les ont affectées toutes les deux ensemble. C'est là une manière possible d'évaluer la proximité linguistique acquise par ces deux lectes au cours du temps – ce que l'on pourrait également appeler leur degré de similarité acquise. Le volow et le mwotlap ont encore un haut degré d'intelligibilité mutuelle, et peuvent être considérés comme deux dialectes d'une même langue. Au passage, dans la mesure où le taux de cohésion mesure le degré de similarité acquise entre deux lectes, il constitue une métrique potentielle pour estimer le degré de proximité entre deux dialectes, d'une manière plus exacte que la notion peu contrôlable d'intelligibilité mutuelle.

Pour tout groupe linguistique, le taux de cohésion mesure à quel point il s'apparente à un groupe généalogique idéal, c-à-d. maximalement solidaire. Dans un arbre théorique tel que la *Figure 2* [§2.1], les groupes ne sont jamais entrecoupés par des innovations transversales, en sorte que le taux de cohésion y est toujours de 100%. Dans le monde réel, ce taux n'est jamais atteint, car l'intersection d'isoglosses constitue la norme plutôt que l'exception.

4.3.4.2 La solidité des groupes généalogiques

Au bout du compte, la robustesse d'un groupe généalogique peut être mesurée de deux manières. En termes absolus, le nombre d'*innovations exclusivement partagées* (ε 'epsilon') correspond au nombre d'attestations effectives de ce groupe en tant que tel. En termes relatifs, le taux de cohésion (k) signale le degré de loyauté interne entre les membres de ce groupe.

Ces deux métriques sont aussi légitimes l'une que l'autre pour évaluer la robustesse d'un groupe généalogique ; et elles sont logiquement indépendantes l'une de l'autre. Un groupe généalogique dont à la fois k et ε seraient élevés aurait une valeur historique évidente : c'est le cas, par exemple, pour la paire *mwotlap–volow*, pour laquelle $k = 92\%$ et $\varepsilon = 14$. À l'inverse, le groupe *vurës–mwesen–mota–nume–mwerlap* présente à la fois une cohésion médiocre ($k = 29\%$) et une attestation faible ($\varepsilon = 2$) : on dira de ce groupe, non pas qu'il n'existe pas (il peut tout à fait s'agir d'un véritable groupe généalogique au sens de la méthode comparative), mais qu'il est moins robuste que d'autres groupes qui le traversent.

Enfin, certains groupes généalogiques sont caractérisés par un taux bas sur l'une des deux dimensions et pas l'autre. Par exemple, le couple *dorig–koro* a une cohésion élevée ($k = 78\%$), mais n'est attesté, dans ma base de données, que $\varepsilon = 5$ fois. De façon symétrique, le groupe des Banks – incluant quinze langues, du lehali au lakon – a une faible cohésion ($k = 30\%$), mais il est confirmé par maintes isoglosses ($\varepsilon = 13$). Ces deux cas de figure suggèrent la possibilité de degrés intermédiaires de solidité, entre les deux extrêmes cités plus haut.

Une solution simple consiste à combiner ces deux métriques en une seule, de manière à estimer la robustesse d'un groupe généalogique. Il suffit alors de multiplier ε par k : le nombre absolu d'innovations partagées, pondéré par le taux de cohésion du groupe. On obtient alors une nouvelle métrique, que je propose d'appeler ***solidité*** du groupe (en anglais *subgroupiness*), noté 'sigma' $\varsigma = \varepsilon \times k$. Ce taux de solidité permet d'évaluer tous les groupes dans une famille donnée, afin de dégager les plus significatifs d'entre eux.

Ainsi, le sous-groupe généalogique formé des six langues des Banks du sud est établi sur la base de $\varepsilon = 7$ innovations exclusivement partagées. Par ailleurs, son taux de cohésion est $k = 0{,}43$: parmi toutes les innovations pertinentes à ce sous-groupe, 43% ont impliqué ces six langues ensemble, venant ainsi confirmer la validité du sous-groupe en question. Le taux de solidité de ce groupe est donc $\varsigma = \varepsilon \times k = 7 \times 0{,}43 = 3{,}00$. Ce dernier chiffre n'est interprétable que relativement aux autres groupes de la région. Sachant que le taux de solidité des groupes de la région a une valeur médiane de $\varsigma = 0{,}3$ (avec un maximum de 12,82 pour le groupe mwotlap–volow), le chiffre de 3,00 est relativement élevé – plus élevé que 90% des sous-groupes.

Le *Tableau 2* fournit les taux de solidité de quelques groupes généalogiques mentionnés dans le présent article.

Tableau 2 — Mesures de cohésion (k) et de solidité (ς) de quelques groupes généalogiques aux Banks–Torres

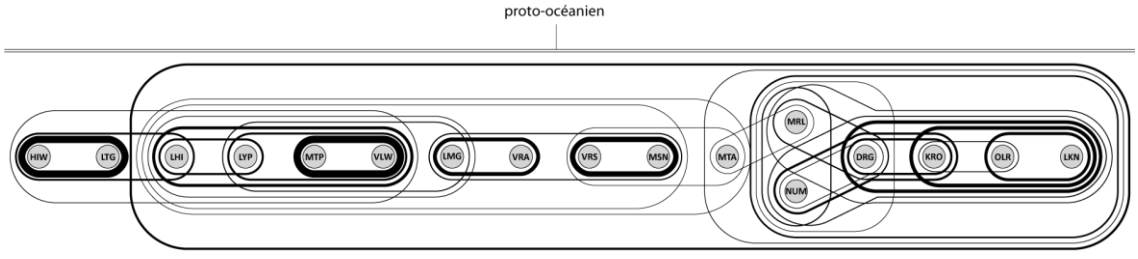
groupe	ε	k	solidité (ς)	
MTP–VLW	14	0,92	ς=14×0,92=	12,82
HIW–LTG	15	0,83	ς=15×0,83=	12,45
LMG–VRA	9	0,75	ς=9×0,75=	6,75
DRG–KRO	5	0,78	ς=5×0,78=	3,90
groupe des îles Banks	13	0,30	ς=13×0,30=	3,90
MRL–NUM–DRG–KRO–OLR–LKN	7	0,43	ς=7×0,43=	3,00
LMG–VRA–VRS–MSN	5	0,44	ς=5×0,44=	2,20
LHI–LYP–VLW–MTP–LMG	3	0,42	ς=3×0,42=	1,26
VRS–MSN–MTA–NUM–MRL	2	0,29	ς=2×0,29=	0,58
VRA–VRS	1	0,57	ς=1×0,57=	0,57

4.3.4.3 Un diagramme glottométrique

Nous avons ainsi calculé les taux de solidité pour les 142 groupes attestés dans la zone Banks–Torres.

Parmi ces 142 groupes, nous avons retenu les plus solides : ceux placés au-dessus d'un seuil de solidité de $\varsigma \geq 1$. Les groupes en dessous de ce seuil sont considérés peu solides, car pourvus d'une solidité inférieure à celle qu'aurait un hapax (groupe supporté par une seule innovation). Pour un seuil de $\varsigma \geq 1$, on obtient 32 groupes considérés significatifs. Ceux-ci sont réunis en une seule figure, que nous appelons **diagramme glottométrique** ou *glottogramme historique* (*Figure 5*) ; la valeur relative de chaque groupe est traduite visuellement par l'épaisseur du trait, exactement proportionnelle au taux de solidité (ς). La fonction de ce type de diagramme est de représenter la généalogie d'une famille linguistique, mieux que ne le ferait une structure arborescente.

Figure 5 — Diagramme glottométrique des langues Banks–Torres



Le diagramme comporte deux parties. Au sommet figure le proto-océanien, ancêtre commun de toutes ces langues. La ligne verticale en haut du diagramme symbolise la dimension temporelle qui sépare l'époque initiale d'unité linguistique et le processus ultérieur de diversification, entre la proto-langue et les langues modernes. La double ligne horizontale – convention empruntée à Ross (1988:9) – indique que les langues descendant de cet ancêtre forment un chaînage de langues, qui ne comporte pas de structure cladistique interne.

Quant à la partie principale de ce diagramme, elle représente la structure généalogique du chaînage linguistique formé par les dix-sept langues du Vanuatu septentrional. De gauche à droite, les langues vont du nord (hiw) au sud (lakon) ; on pourrait en principe étendre ce diagramme en ajoutant davantage de langues parlées au sud des îles Banks (cf. François *et al.* 2015:16).

Faute de place, je me contenterai ici de quelques commentaires essentiels. Tout d'abord, les taux de solidité, ainsi que le diagramme qui en est dérivé, confirment bien que les langues du nord du Vanuatu forment un chaînage dans lequel les groupes généalogiques se chevauchent régulièrement. Ainsi, on voit que le lemerig LMG forme l'intersection d'au moins deux groupes : l'un côté nord [LHI, LYP, MTP, VLW, *LMG*] (ς = 1,26), l'autre côté sud [*LMG*, VRA, VRS, MSN] (ς = 2,20)[12]. De la même manière, le mota MTA forme un chaînon transitionnel entre le groupe des Banks du nord–centre [LHI…MTA] (ς = 1,03) et celui des Banks du centre–sud [MTA…LKN] (ς = 1,30).

Le groupe des *Banks du sud* proprement dit [MRL…LKN] est l'illustration parfaite d'une chaîne dialectale, où chaque paire de langues géographiquement adjacentes forme une unité généalogique : [MRL, NUM], [NUM, DRG], [MRL, DRG], [DRG, KRO], [KRO, OLR], [OLR, LKN]… Ce schéma en chaînage se retrouve même avec les triplets voire les quadruplets de langue, dont la concaténation force le respect : [NUM, DRG, MRL], [NUM, DRG, KRO], [KRO, OLR, LKN], [DRG, KRO, OLR, LKN], [NUM, DRG, KRO, OLR, LKN], [MRL, DRG, KRO, OLR, LKN].

4.3.4.4 Glottométrie et géographie linguistique

Les résultats de la glottométrie historique présentent un fort potentiel pour reconstruire la structure des réseaux anthropologiques à date ancienne. Ils peuvent également être projetés sur une carte géographique, permettant de visualiser dans l'espace les relations généalogiques entre les divers membres d'un chaînage (*Figure 6*).

Dans l'ensemble, la concaténation des langues est cohérente avec leur distribution spatiale. Ceci reflète l'intuition : les membres d'un continuum linguistique ont tendance à partager beaucoup d'innovations avec leurs voisins immédiats, et moins avec leurs voisins plus éloignés – comme on peut s'y attendre. Mais il s'agit en réalité là d'un résultat empirique, moins trivial qu'il n'y paraît. Ainsi, le fait que la distribution des isoglosses historiques – dont certaines remontent à trois mille ans – soit généralement cohérente avec la distribution spatiale des langues modernes implique une remarquable stabilité dans l'ancrage géographique de ces communautés linguistiques. Les résultats glottométriques suggèrent fortement que ces dernières se sont développées *in situ* au cours des trois derniers millénaires, sans déplacements majeurs de population[13].

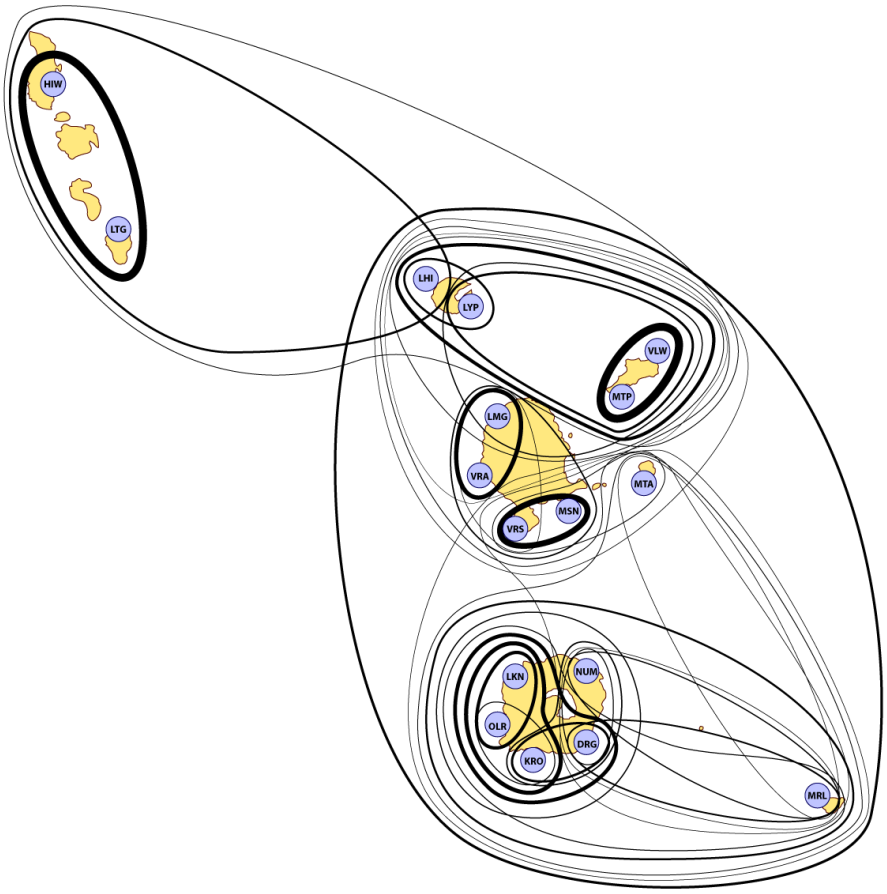
Au bout du compte, une observation attentive de la carte montre que les distances linguistiques ne se contentent pas de refléter la distance géographique. Par exemple, alors que le mwotlap est géographiquement plus proche du mota que du löyöp, la position de ces trois langues dans la topologie glottométrique (*Figure 5*) révèle que le mwotlap et le mota sont généalogiquement assez éloignés (k = 35%). Ceci est égale-

[12] Entre autres innovations diagnostiques, le premier de ces deux groupes est défini par le changement **wa$^{\eta}$ga* > **o$^{\eta}$ga* 'bateau' (cf. *Tableau 1*); le second par la dissimilation **mama^{n}ri^{n}ri* > **mamaɣi^{n}ri* 'froid' (François 2015:846).

[13] François (2011b:181-182) arrive à une conclusion similaire à l'échelle de l'archipel du Vanuatu plus globalement, en observant la distribution du protophonème *R.

ment visible sur la carte, où MTP et MTA ne sont liées entre elles par aucune isoglosse (sauf à plus grande échelle, celle des îles Banks dans leur ensemble). Généalogiquement parlant, le mwotlap est nettement tourné vers le nord-ouest, et le mota vers le sud. Manifestement, la société ancienne de Mota entretenait des relations sociales (commerce, mariage) bien moins intenses avec l'île de Motalava voisine qu'avec les îles situées au sud-ouest (Vanua Lava) ou au sud (Gaua) ; et même l'île de Merelava, pourtant bien plus éloignée que Motalava, était pour Mota un partenaire proche.

Figure 6 — Carte glottométrique des langues Torres–Banks



C'est d'ailleurs ce type de configuration qui pourrait être interprétée, dans certains cas, comme l'indice d'une possible migration. Au vu de la grande proximité généalogique – manifeste sur la carte – entre le mota et les langues de Gaua ou de Merelava, on pourrait suggérer que la langue mota se soit développée historiquement plutôt dans la zone sud des îles Banks, par exemple dans des villages sur la côte est de Gaua, avant de migrer plus récemment vers l'île de Mota plus au nord. Dans le cas particulier des îles Banks, je ne pense pas qu'une telle hypothèse soit indispensable : on peut également imaginer que la langue ait été depuis toujours ancrée dans l'île de Mota, et que ce soit seulement le jeu des alliances et des mariages interinsulaires – favorisé par de bonnes techniques de navigation maritime – qui ait imprimé au réseau glottométrique cette forte inclination vers le sud. Mais dans d'autres familles de langues, une telle distortion entre généalogie et géographie pourrait tout à fait s'expliquer par des mouvements de population.

4.3.5 BILAN ET DISCUSSION

La glottométrie historique propose une manière inédite de reconstruire et représenter les liens de parenté entre les langues. Il est normal que cette approche innovante soulève de nouvelles questions.

Ainsi, les pages précédentes ont proposé trois manières différentes de représenter les résultats de ses calculs : sous la forme d'un tableau de nombres (*Tableau 2* p.24), d'un diagramme (*Figure 5*), ou d'une carte (*Figure 6*). Pour peu que la linguistique historique s'empare de notre modèle, il est probable que de nouvelles idées seront avancées afin de rendre les résultats glottométriques encore plus clairs. On pourrait améliorer la citabilité de ces résultats ; automatiser la production de glottogrammes et de cartes ; mettre à profit les couleurs (à l'instar des cartes de dialectométrie [§4.1]) ; inventer des représentations tri-dimensionnelles ; développer des cartes électroniques interactives… On peut également imaginer que les principes théoriques sous-jacents à notre approche inspirent de nouveaux raffinements au modèle glottométrique lui-même, donnant lieu à de nouveaux calculs pour compléter les mesures de *cohésion* et de *solidité*. Le lecteur est convié à améliorer ces premières propositions.

Enfin, je terminerai cette étude par une brève discussion de deux questions : celle de la représentation de la chronologie ; et celle de la nature des proto-langues que l'on peut reconstruire.

4.3.5.1 Peut-on ordonner les groupes dans le temps ?

Une question lancinante est celle de la chronologie des innovations : comment la glottométrie peut-elle intégrer la dimension temporelle ? Le modèle de l'arbre avait au moins cet avantage, qu'il permettait une lecture chronologique de ses ramifications, en descendant d'un nœud à l'autre ; ne pourrait-on faire de même avec des groupes généalogiques entrecroisés ? Une première indication temporelle apparaît en haut du diagramme, reliant la proto-langue au chaînage moderne ; mais ne peut-on aller plus loin ?

On pourrait imaginer, par exemple, que les groupes les plus larges précèdent dans le temps les groupes les plus réduits. Dans une première phase, le vaste groupe des îles Banks aurait connu toutes les innovations qui le distinguent du proto-océanien ; ensuite, ce groupe des Banks se serait ensuite scindé en deux groupes (centre-nord-Banks et centre-sud-Banks, avec le mota comme maillon commun) ; et ainsi de suite. C'est à ce genre de logique que le raisonnement cladistique nous a habitués, dans l'idée que les langues se diversifieraient principalement par scissions successives [§2.2]. Dans cet état d'esprit, on pourrait concevoir un glottogramme à plusieurs étages, correspondant à plusieurs phases historiques : l'ancêtre commun (ex. le proto-océanien, en haut de la figure) se diviserait en deux groupes, puis en quatre…, jusqu'à obtenir le chaînage le plus complexe.

L'ennui, c'est que le plus souvent, il n'est pas possible de connaître à coup sûr l'ordre dans lequel les innovations se sont produites ; en sorte qu'insérer la chronologie dans nos représentations ne sera souvent que spéculatif. Dans une minorité de cas (20% dans ma base de données du nord du Vanuatu), il est possible de reconstituer une chronologie relative entre les innovations. Mais cette chronologie ne va pas toujours du plus grand au plus petit : j'ai pu ainsi montrer (François 2011a:201) que certaines innovations ciblant le groupe [KRO-OLR-LKN] sont nécessairement antérieures à d'autres

innovations qui ont affecté tout le nord du Vanuatu[14]. De tels contre-exemples nous montrent qu'on ne saurait décider *a priori* de l'ordre chronologique des innovations : ce dernier ne peut être examiné qu'au cas par cas, et la plupart du temps – en l'absence de données empiriques fiables – devra rester indéterminé.

En somme, alors que les axiomes du modèle cladistique imposent artificiellement un ordre chronologique *a priori*, l'approche glottométrique se doit de rester agnostique concernant la chronologie relative entre groupes généalogiques.

4.3.5.2 Quelles proto-langues dans un modèle diffusionniste ?

Ces observations nous incitent également à la prudence lorsqu'il s'agit de reconstituer les proto-langues. Le proto-Banks, par exemple, peut-il être reconstruit ? A-t-il seulement existé – ou bien n'est-il qu'une illusion ?

La réponse dépendra de ce que l'on entend par « proto-langue ». Le genre de proto-langue que le modèle de l'arbre nous incite à reconstruire (cf. Pulgram 1961) est un système qui se veut lisse, homogène, dépourvu de variation dialectale interne. Si l'on reprend l'arbre de la *Figure 2* [§2.1], on reconstruira ainsi le proto-MNO comme une langue dans laquelle ni N, ni O ne se seront pas encore distinguées de M. Le système de pMNO sera simplement celui reconstruit pour pKLMNO, augmenté de toutes les innovations communes à M N et O, et seulement celles-ci. Le modèle implique que ces dernières innovations doivent être chronologiquement antérieures à celles de N, de O ou de pNO. Une fois de plus, de tels axiomes ont l'avantage de rendre le modèle cladistique simple, prédictible, élégant… mais avec l'inconvénient qu'ils imposent ainsi une chronologie factice, artefact du modèle choisi plutôt qu'analyse dérivable des faits empiriques. En outre, une telle proto-langue idéalisée est peu crédible, car elle ne ressemble pas aux langues vivantes telles qu'on les observe sur le terrain, et qui sont marquées par la variation et la diversité interne.

Mais le concept de proto-langue redevient plausible si on le conçoit comme un ensemble de dialectes, certes mutuellement intelligibles, mais susceptibles d'avoir déjà commencé à se différencier les uns des autres. Ce processus de différenciation interne peut avoir commencé très tôt, et se poursuivre durant des siècles, tout au long de l'existence de cette proto-langue – jusqu'à ce que cette dernière finisse par succomber à sa fragmentation interne. Pour reprendre l'arbre de la *Figure 2*, il faut être prêt à penser une proto-langue pMNO dans laquelle, éventuellement, certaines innovations propres à O seul, ou à N et O ensemble, se seraient déjà produites, au moment où interviendraient de nouvelles innovations propres à pMNO dans son ensemble [§4.3.5.1]. Si la pensée cladistique nous empêche de concevoir une telle configuration, celle-ci devient parfaitement compréhensible si l'on adopte un parti diffusionniste : on dira simplement que l'ensemble MNO se caractérise par une relative unité linguistique (M N et O étant mutuellement intelligibles), mais qu'il comporte également une diversité dialectale, due à certaines innovations qui auront déjà eu lieu en son sein.

Une proto-langue doit donc être conçue potentiellement comme un système hétérogène, une sorte de *diasystème* (Weinreich 1954:390), dénominateur commun de tous les dialectes qui la composent. Appliquons ce raisonnement au nord du Vanuatu, en nous référant à la liste d'innovations donnée dans le *Tableau 1* [§4.3.2]. Si l'on souhaite reconstruire le proto-Banks, il faudra bien sûr y localiser, à un moment ou à un autre de

[14] Voir la section §3.4, et la citation de Chambon (2011).

son histoire, l'innovation lexicale n°1 (le remplacement de **panako* par **^{m}balu* pour 'voler'). Mais il est fort possible que ce même proto-Banks, au moment de la diffusion de cette innovation n°1, soit déjà caractérisé par une variation dialectale interne dans le nom du 'bateau' (certains dialectes reflétant **o^{ŋ}ga* plutôt que **wa^{ŋ}ga* : innov. n°4) ; dans le nom de l'écrevisse (n°5) ; dans la forme du verbe 'être allongé' (n°7) ou celle du pronom de 3^{e} personne triel (n°10)…

En l'absence de connaissance sur l'ordre relatif entre les innovations, rien ne permet de savoir si les isoglosses réduites précèdent ou suivent celles de plus grande étendue. C'est aussi ce type de configuration-là que représente le glottogramme de la *Figure 5* : par sa structure, il est délibérément agnostique en ce qui concerne la chronologie des innovations. On pourrait d'ailleurs arguer que le travail de la méthode comparative n'est pas tant de reconstruire des *proto-langues* en tant que systèmes entiers, mais plutôt de reconstituer des innovations individuelles – le plus souvent, sans pouvoir se prononcer quant à leur séquence et leur datation relative.

Pendant ses siècles d'évolution, le proto-Banks aura donc vu s'accumuler en son sein des innovations qui tendaient à le fragmenter ; mais aussi, en parallèle, d'autres innovations qui continuaient à se propager d'un bout à l'autre de ce petit archipel, et en renforçaient la cohésion. Pendant un temps, il a pu s'agir d'un mouvement pendulaire (cf. les « changements en accordéon » de Chambon), alternant divergence et convergence (François 2011a). Puis, à mesure que les vastes ensembles accumulaient les divergences internes, les zones d'intelligibilité mutuelle se réduisaient. Petit à petit, au fil des siècles, les membres de cette communauté initialement homogène – celle des îles Torres et Banks – ont fini par former la mosaïque linguistique que nous connaissons aujourd'hui, dans laquelle l'intelligibilité mutuelle n'existe qu'entre langues immédiatement adjacentes, et ne va guère au-delà.

En somme, même dans une approche glottométrique, il demeure légitime de parler de proto-langues, et de formuler des hypothèses de reconstructions – dans la lignée de la méthode comparative. Simplement, si l'on souhaite vraiment adopter une démarche réaliste visant à reconstruire l'histoire effective des familles de langues, il est nécessaire d'affiner nos outils théoriques. Malgré la difficulté que cela représente, nous devons être prêts à concevoir des proto-langues non plus comme une abstraction idéalisée, mais comme des réseaux d'idiolectes marqués par l'hétérogénéité interne, la variation inter-individuelle, la synchronie dynamique… – des proto-langues aussi riches et complexes que les langues vivantes que nous connaissons.

5 CONCLUSION

Contrairement à une croyance bien ancrée en linguistique historique, il n'y a aucune raison de penser que la diversification des langues doive prendre la forme d'un arbre généalogique, doté de branchements binaires et définitifs.

En effet, il faut garder à l'esprit que la structure généalogique interne de toute famille de langues repose sur un ensemble d'innovations linguistiques qui se sont diffusées d'idiolecte à idiolecte, le long d'un réseau social formant un continuum. Sachant que ces innovations sont en principe indépendantes les unes des autres, les isoglosses qu'elles définissent peuvent parfaitement se chevaucher. Mis à part quelques cas exceptionnels de scission définitive au sein d'une communauté, la situation normale – celle qui devrait constituer la norme en phylogénétique – est celle où les groupes généalogiques sont enchevêtrés les uns dans les autres.

De telles familles constituent des *chaînages linguistiques*, qui ne peuvent être analysés et représentés qu'à l'aide d'un modèle non cladistique. La *Wellentheorie*, avec ses « ondes » superposées au cours du temps, fournit l'intuition centrale dont nous avions besoin pour proposer un contre-modèle de la généalogie des langues. Mais il fallait également trouver un moyen de formaliser ces intuitions, à l'aide d'une méthode fiable et falsifiable. Parmi les diverses réponses possibles à ce défi, la Glottométrie historique est une approche empirique et quantitative, qui vise à dégager avec précision la structure généalogique des familles linguistiques. Cette approche diffusionniste de la phylogénétique nous fournit des outils théoriques et pratiques pour reconstituer l'histoire des familles linguistiques d'une manière précise et réaliste, tout en demeurant fidèles au génie de la méthode comparative.

Alexandre FRANÇOIS
CNRS–LACITO (Langues et Civilisations à Tradition Orale)
Australian National University

6 BIBLIOGRAPHIE

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

ALLIÈRES, Jacques. 2001. *Manuel de linguistique romane*. Bibliothèque de grammaire et de linguistique, 10. Paris : Honoré Champion.

ANTTILA, Raimo. 1989. *Historical and Comparative Linguistics*. Amsterdam-Philadelphia: Benjamins.

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

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

BLUST, Robert. 2013. *Austronesian Comparative Dictionary*, édition sur internet. [http://www.trussel2.com/ACD, accessed 22 May 2013].

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

BOWERN, Claire. 2006. Another look at Australia as a linguistic area. In Y. Matras, A. McMahon & N. Vincent (eds.), *Linguistic Areas*. Basingstoke: Palgrave Macmillan, 244–265.

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

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

BRYANT, David, Flavia FILIMON, & Russell D. GRAY. 2005. Untangling our past: Languages, Trees, Splits and Networks. In R. Mace, C.J. Holden & S. Shennan (eds.), *The Evolution of Cultural Diversity: Phylogenetic Approaches*. London: UCL Press, 67-83.

CAMPBELL, Lyle. 2004. *Historical linguistics: An introduction*. Edinburgh: Edinburgh University Press.

CAMPBELL, Lyle, & William J. POSER. 2008. *Language classification: History and method*. Cambridge: Cambridge University Press.

CAVALLI-SFORZA, Luigi Luca, & Marcus W. FELDMAN. 1981. *Cultural transmission and evolution: a quantitative approach*. Vol. 16. Princeton: Princeton University Press.

CHAMBERS, Jack K., & Peter TRUDGILL. 1998. *Dialectology*. Cambridge textbooks in linguistics. Cambridge: Cambridge University Press.

CHAMBON, Jean-Pierre. 2011. Note sur la diachronie du vocalisme accentué en istriote/istroroman et sur la place de ce groupe de parlers au sein de la branche romane. *Bulletin de la Société de Linguistique de Paris* 106.1: 293-303.

CHAPPELL, Hilary. 2001. Language contact and areal diffusion in Sinitic languages: Problems for typology and genetic affiliation. In Aikhenvald & Dixon, 328–357.

CROFT, William. 2000. *Explaining Language Change: An Evolutionary Approach*. Longman Linguistics Library. London: Longman.

CURRIE, Thomas E. ; Simon J. GREENHILL, & Ruth MACE. 2010. Is horizontal transmission really a problem for phylogenetic comparative methods? A simulation study using continuous cultural traits. *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 365.1559: 3903-3912.

DIXON, R.M.W. 1997. *The Rise and Fall of Languages*. Cambridge: Cambridge University Press.

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

DUNN, Michael. 2014. Language phylogenies. In Claire Bowern & Bethwyn Evans (eds), *The Routledge Handbook of Historical Linguistics*. New York: Routledge, 190-211.

DUNN, Michael, Stephen LEVINSON, Eva LINDSTRÖM, Ger REESING, & Angela TERRILL. 2008. Structural phylogeny in historical linguistics: Methodological explorations applied in Island Melanesia. *Language* 84 (4):710-759.

DURIE, Mark & Malcolm ROSS (eds). 1996. *The Comparative Method Reviewed: Regularity and irregularity in language change*. Oxford: Oxford University Press.

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

ENFIELD, Nicholas. 2008. Transmission biases in linguistic epidemiology. *Journal of Language Contact* Thema 2:299-310.

ERNST, Gerhard; Martin-Dietrich GLEßGEN, Christian SCHMITT & Wolfgang SCHWEICKARD (eds.) 2003-2008. *Romanische Sprachgeschichte — Ein Internationales Handbuch zur Geschichte der Romanischen Sprachen/Histoire linguistique de la Romania — Manuel international d'histoire linguistique de la Romania*. 3 volumes. Berlin: De Gruyter Mouton.

FORSTER, Peter; Alfred TOTH & Hans-Jürgen BANDELT. 1998. Evolutionary network analysis of word lists: Visualising the relationships between Alpine Romance languages. *Journal of Quantitative Linguistics* 5 (3):174-187.

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

FRANÇOIS, Alexandre. 2001. Contraintes de structures et liberté dans l'organisation du discours: Une description du mwotlap, langue océanienne du Vanuatu. Thèse de doctorat en linguistique. Paris: Université Paris-IV Sorbonne. 1078 pp.

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

FRANÇOIS, Alexandre. 2007. Noun articles in Torres and Banks languages: Conservation and innovation. In Jeff Siegel, John Lynch & Diana Eades (eds.), *Language Description, History and Development: Linguistic indulgence in memory of Terry Crowley*. Creole Language Library 30. New York: Benjamins, 313-326.

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. 2013. Shadows of bygone lives: The histories of spiritual words in northern Vanuatu. In R. Mailhammer (ed.), *Lexical and structural etymology: Beyond word histories*. Studies in Language Change. Berlin: De Gruyter Mouton, 185-244.

FRANÇOIS, Alexandre. 2014. Trees, Waves and Linkages: Models of Language Diversification. In Claire Bowern & Bethwyn Evans (eds), *The Routledge Handbook of Historical Linguistics*. New York: Routledge, 161-189.

FRANÇOIS, Alexandre. 2015. Temperature terms in northern Vanuatu. In Maria Koptjevskaja Tamm (ed.), *The Linguistics of Temperature*. Amsterdam, New York: John Benjamins, 832-857.

FRANÇOIS, Alexandre. ss presse. The history of personal pronouns in northern Vanuatu. In Konstantin Pozdniakov (ed.), *Reconstruction et classification généalogique : tendances actuelles*. *Faits de Langues*. Bern: Peter Lang.

FRANÇOIS, Alexandre; Michael FRANJIEH; Sébastien LACRAMPE; Stefan SCHNELL. 2015. The exceptional linguistic density of Vanuatu. In A. François; S. Lacrampe; S. Schnell & M. Franjieh (eds), *The Languages of Vanuatu: Unity and Diversity*. Studies in the Languages of Island Melanesia. Canberra: Asia–Pacific Linguistics Open Access, 1–21.

GARRETT, Andrew. 2006. Convergence in the formation of Indo-European subgroups: Phylogeny and chronology. In P. Forster & C. Renfrew (eds.), *Phylogenetic methods and the prehistory of languages*. Cambridge: McDonald Institute for Archaeological Research, 139-151.

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

GILES, Howard, & Tania OGAY. 2007. Communication accommodation theory. In B. B. Whaley & W. Samter (eds.), *Explaining communication: Contemporary theories and exemplars*. London: Routledge, 293-310.

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

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

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

GRAY, Russell D., Alexei J. Drummond, & Simon J. Greenhill. 2009. Language Phylogenies Reveal Expansion Pulses and Pauses in Pacific Settlement. *Science* 323 (5913): 479-483.

GREENHILL, Simon J., & Russell D. GRAY. 2009. Austronesian language phylogenies: myths and misconceptions about Bayesian computational methods. In A. Adelaar & A. Pawley (eds.), *Austronesian historical linguistics and culture history: a festschrift for Robert Blust*. Pacific Linguistics, 601. Canberra: Australian National University, 375-397.

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

GUARISMA, Gladys, & Wilhelm J.G. MÖHLIG (eds.). 1986. *La méthode dialectométrique appliquée aux langues africaines*. Berlin: Reimer.

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

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

HASPELMATH, Martin, & Uri TADMOR. 2009. *Loanwords in the World's Languages: A comparative handbook*. Berlin: Mouton de Gruyter.

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*. Trends in Linguistics: Studies and Monographs, 34. Berlin: de Gruyter.

HOLTON, Gary. 2011. A Geo-linguistic Approach to Understanding Relationships within the Athabaskan Family. Paper read at the international workshop *Language in Space: Geographic Perspectives on Language Diversity and Diachrony*. Boulder, Colorado.

HUEHNERGARD, John, & Aaron RUBIN. 2011. Phyla and Waves: Models of Classification. Semitic Languages. In S. Weninger, G. Khan, M. P. Streck & J. Watson (eds.), *Semitic Languages: An International Handbook*. Handbücher zur Sprach- und Kommunikationswissenschaft, 36. Berlin: de Gruyter Mouton, 259-278.

KALYAN, Siva, & Alexandre FRANÇOIS. f/c. Freeing the Comparative Method from the Tree Model: A framework for Historical Glottometry. In R. Kikusawa & L. Reid (eds.), *Let's talk about trees: Tackling Problems in Representing Phylogenic Relationships among Languages*. Senri Ethnological Studies. Osaka: National Museum of Ethnology.

KORN, Agnes. 2003. Balochi and the concept of North-West Iranian. In Carina Jahani & Agnes Korn (eds.), *The Baloch and their neighbours: Ethnic and linguistic contact in Balochistan in historical and modern times*. Wiesbaden: Reichert, 49-60.

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

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

LABOV, William. 1994. *Principles of linguistic change: Internal factors*. Oxford: Blackwell.

LABOV, William. 2001. *Principles of linguistic change: Social factors*. Oxford: Blackwell.

LABOV, William. 2007. Transmission and diffusion. *Language* 83 (2):344-387.

LECOINTRE, Guillaume & Hervé LE GUYADER. 2001. *Classification phylogénétique du vivant*. Paris: Belin.

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

LICHTENBERK, Frantisek. 2013. Development of reason and cause markers in Oceanic. *Oceanic Linguistics* 52.1 (June 2013): 86-105.

LYNCH, John. 2000. Linguistic subgrouping in Vanuatu and New Caledonia. In B. Palmer & P. A. Geraghty (eds.), *Proceedings of the Second International Conference on Oceanic Linguistics (SICOL), vol. 2: Historical and descriptive studies*. Pacific Linguistics, 505. Canberra: University of the South Pacific, 155-184.

MILROY, Lesley. 1987. *Language and social networks*. Language in Society. Oxford: Blackwell.

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

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. In E. Zeitoun & P. J.-K. Li (eds.), *Selected Papers from the Eighth International Conference on Austronesian Linguistics*. Symposium Series of the Institute of Linguistics: Academia Sinica, 95-138.

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

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

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

RAMAT, Paolo. 1998. The Germanic languages. In A. G. Ramat & P. Ramat (eds.), *The Indo-European languages*. London: Routledge, 380-414.

RANKIN, Robert. 2003. The Comparative Method. In B. D. Joseph & R. D. Janda (eds.), *The Handbook of Historical Linguistics*. Oxford: Blackwell, 183-212.

ROSS, Malcolm. 1988. *Proto-Oceanic and the Austronesian languages of Western Melanesia*. Pacific Linguistics. Canberra: Australian National University.

ROSS, Malcolm. 1996. Contact-induced change and the Comparative Method: Cases from Papua New Guinea. In Durie & Ross (eds), 180-217.

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

ROSS, Malcolm. 2001. Contact-induced change in Oceanic languages in North-West Melanesia. In Aikhenvald & Dixon, 134–166.

SAUSSURE, Ferdinand de. 1985 [1916]. *Cours de linguistique générale*. Édition critique de Tullio de Mauro. Paris: Payot.

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

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

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

SCHUCHARDT, Hugo. 1900 [1870]. *Über die Klassifikation der romanischen Mundarten: Probevorlesung gehalten zu Leipzig am 30. april 1870*. Graz: Styria.

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

STREET, Richard L., & Howard Giles. 1982. Speech accommodation theory: A social cognitive approach to language and speech behavior. In M. E. Roloff & C. R. Berger (eds.), *Social cognition and communication*. Beverly Hills: Sage, 193-226.

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

TOULMIN, Matthew. 2009. *From linguistic to sociolinguistic reconstruction: the Kamta historical subgroup of Indo-Aryan*. Studies in Language Change, 604. Canberra: Pacific Linguistics.

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

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

WANG, William S-Y. & James W. MINETT. 2005. Vertical and horizontal transmission in language evolution. *Transactions of the Philological Society* 103.2: 121-146.

WEINREICH, Uriel. 1954. Is a structural dialectology possible? *Word* 10.2-3 (1954): 388-400.

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