

---

## Gènes et Langues : une longue histoire commune ?

*Genes and Languages: a long and common history?*

**Mahé Ben Hamed et Pierre Darlu**

---



### Édition électronique

URL : <https://journals.openedition.org/bmsap/5363>

DOI : 10.4000/bmsap.5363

ISSN : 1777-5469

### Éditeur

Société d'Anthropologie de Paris

### Édition imprimée

Date de publication : 1 décembre 2007

Pagination : 243-264

ISSN : 0037-8984

### Référence électronique

Mahé Ben Hamed et Pierre Darlu, « Gènes et Langues : une longue histoire commune ? », *Bulletins et mémoires de la Société d'Anthropologie de Paris* [En ligne], 19 (3-4) | 2007, mis en ligne le 03 février 2011, consulté le 01 juin 2021. URL : <http://journals.openedition.org/bmsap/5363> ; DOI : <https://doi.org/10.4000/bmsap.5363>

---

Ce document a été généré automatiquement le 1 juin 2021.



Les contenus des *Bulletins et mémoires de la Société d'Anthropologie de Paris* sont mis à disposition selon les termes de la licence Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

---

# Gènes et Langues : une longue histoire commune ?<sup>1</sup>

*Genes and Languages: a long and common history?*

Mahé Ben Hamed et Pierre Darlu

---

## Introduction

- 1 L'intérêt des biologistes pour le discours des linguistes est une constante que l'on peut observer depuis le milieu du XIXe s. jusqu'à nos jours. Sans doute les naturalistes espèrent-ils contribuer à la connaissance des langues, de leur diversité et de leurs origines grâce à l'originalité de leurs réflexions et de leurs méthodes. Un inventaire reste cependant à faire, car l'originalité n'a pas toujours été du côté des naturalistes. De plus, l'ingénuité avec laquelle les biologistes abordent une science aussi complexe que la linguistique provoque souvent sourires amusés ou critiques acerbes chez les linguistes...
- 2 Dans cette ébauche d'analyse, nous tenterons d'éclairer le débat en nous appuyant sur quelques textes historiques. Nous examinerons ensuite la façon dont les linguistes abordent les approches phylogénétiques maintenant courantes en biologie. Enfin, nous décrirons comment les généticiens intègrent un certain type d'informations linguistiques dans leur recherche sur la dynamique des populations (et sur l'origine de l'homme).

## Un peu d'histoire

### De Darwin aux lois de Mendel

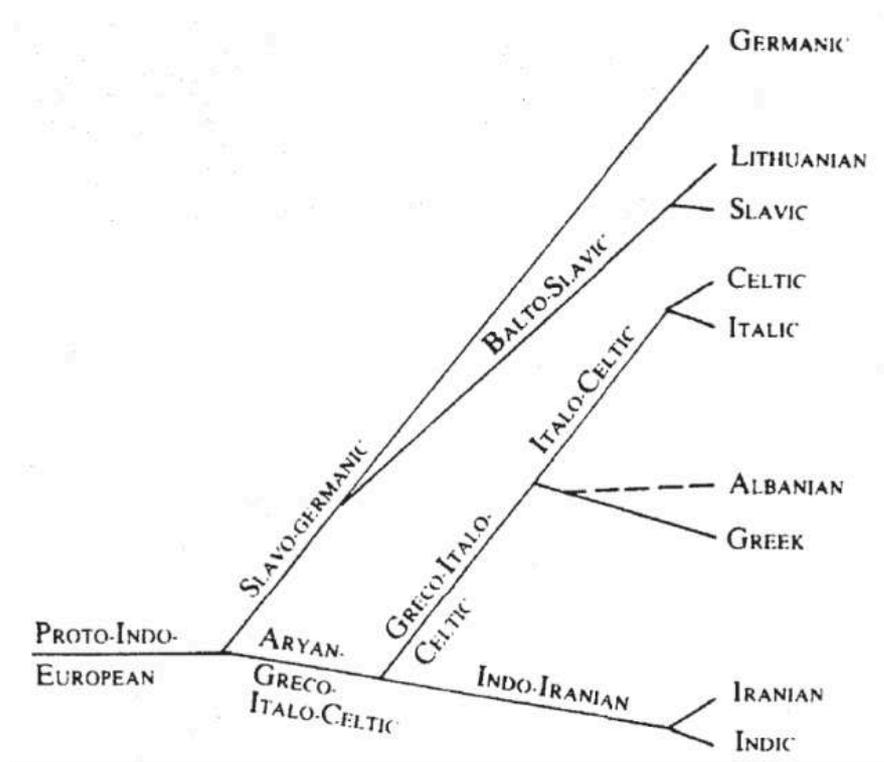
- 3 Au milieu du XIX<sup>e</sup> s., dans son *Origine des espèces*, Charles Darwin (Darwin 1859a, b) prend l'exemple des langues pour expliquer le principe de l'évolution biologique et sa représentation sous la forme d'un arbre généalogique :

- 4 « Pour mieux faire comprendre cet exposé de la classification, prenons un exemple tiré des diverses langues humaines. Si nous possédions l'arbre généalogique complet de l'humanité, un arrangement généalogique des races humaines présenterait la meilleure classification des diverses langues parlées actuellement dans le monde entier ; [...] Les divers degrés de différences entre les langues dérivant d'une même souche devraient donc s'exprimer par des groupes subordonnés à d'autres groupes ; mais le seul arrangement convenable ou même possible serait encore d'ordre généalogique. Ce serait en même temps l'ordre strictement naturel, car il rapprocherait toutes les langues mortes et vivantes, suivant leurs affinités les plus étroites, en indiquant la filiation et l'origine de chacune d'elles (traduction de la 6<sup>e</sup> édition anglaise par E. Barbier 1880, Reinwald & C<sup>ie</sup> Ed.).
- 5 La position de Darwin est particulièrement instructive et très actuelle. Pour lui, l'arbre évolutif des populations offre la possibilité de retracer l'évolution naturelle des langues qu'elles parlent. Cette position est exactement celle qui est tenue dans plusieurs travaux récents de génétique des populations, comme nous le verrons plus loin.
- 6 Dans *La Descendance de l'Homme* (Darwin 1871), Darwin revient sur cette idée de relations naturelles et généalogiques entre les langues, en reprenant à son compte « *l'intéressant parallélisme entre le développement des espèces et celui des langages* » proposé par Sir C. Lyell dès 1863.
- 7 « Il est à remarquer, et c'est un fait extrêmement curieux, que les causes qui expliquent la formation des langues différentes expliquent aussi la formation des espèces distinctes ; [...] Nous rencontrons, dans les langues distinctes, des homologies frappantes dues à la communauté de descendance, et des analogies dues à un procédé semblable de formation.[...] Les langues, comme les êtres organisés peuvent se classer en groupes subordonnés ; on peut aussi les classer naturellement selon leur dérivation. [...] Si deux langages contiennent un grand nombre de mots et de formes de construction identique, on est d'accord pour reconnaître qu'ils dérivent d'une source commune, quand bien même ils pourraient différer beaucoup par quelques autres points. » (traduction de la 2<sup>e</sup> édition anglaise par E. Barbier 1891, Reinwald & C<sup>ie</sup> Ed.)
- 8 Darwin insiste donc sur le parallélisme entre les mécanismes de différenciation des langues et ceux qui prévalent pour expliquer l'évolution des caractères biologiques. Des espèces qui possèdent les mêmes caractères issus d'un ancêtre commun font, de ce fait, la démonstration de leur parenté biologique. Il en est de même pour les langues. En termes actuels, nous dirions que le partage de caractères dérivés permet d'inférer la généalogie des espèces aussi bien que celle des langues. C'est le fondement des méthodes cladistiques qui seront évoquées plus loin.
- 9 On peut se demander si les idées de Darwin sur l'évolution des langues avaient une réelle originalité aux yeux des linguistes de l'époque ou si, au contraire, elles n'étaient qu'un tissu de banalités intelligemment mises en valeur. Des études plus détaillées sur l'histoire des idées qui circulaient dans la première moitié du XIX<sup>e</sup> s. permettraient de répondre (Tort 1980). Cependant, on peut noter qu'Auguste Schleicher, sans conteste « un évolutionniste pré-darwinien », introduisit la notion d'arbre généalogique des langues dès 1853 (sur des langues slaves et lithuaniennes) bien avant la publication de *L'Origine des Espèces* (1859) [voir *fig. 1* pour un arbre des langues indo-européennes de Schleicher (Schleicher 1853)]. Sans doute sa formation initiale de naturaliste et de botaniste, vantée par Ernest Haeckel dans son *Anthropogénie*, l'y a-t-il préparé, ce qui expliquerait également pourquoi il appelle les naturalistes à prendre

davantage en considération les problèmes linguistiques, en espérant qu'ils proposent leur propre solution. Dans sa lettre publique à Haeckel, datée de 1863, Schleicher écrit en effet ceci : « *En m'adressant à toi [...], je parle surtout aux naturalistes, que je voudrais voir plus instruits dans la science des langues qu'ils ne l'ont été jusqu'ici. Et je n'entends pas seulement par là l'analyse physiologique des sons vocaux,[...] mais aussi la préoccupation des différences linguistiques et de leur importance pour l'histoire naturelle du genre Homme[...]. De même un de mes vœux les plus chers, c'est que la méthode des sciences naturelles trouve de plus en plus de faveur auprès des linguistes.* »

Fig. 1 - Représentation arborescente des filiations entre langues indo-européennes (Schleicher 1863).

Fig. 1—Tree representation of relationships between Indo-European languages (Schleicher 1863).

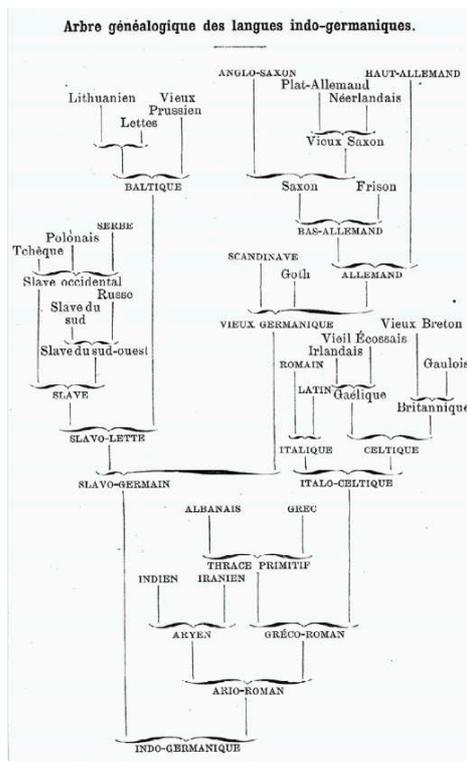


- 10 Nous sommes ainsi devant une situation plutôt paradoxale : un linguiste, parmi les plus engagés dans une explication évolutive de la diversité des langues, faisant appel aux naturalistes pour l'épauler dans son engagement, et des naturalistes profitant des évidences sur l'évolution linguistique pour défendre leurs propres idées sur l'évolution du monde vivant.
- 11 Dans son livre *Anthropogénie ou Histoire de l'Évolution humaine* (Haeckel 1874), Haeckel reprend les mêmes idées que celles développées par Darwin, parfois presque mot à mot :
- 12 « C'est qu'en présence des merveilleux résultats obtenus par la linguistique comparée, il faut vraiment se boucher les yeux pour ne pas voir l'évolution naturelle du langage. Pour un naturaliste, cela va de soi. »

- 13 Puis, reprenant l'exemple des langues « indo-germaniques » proposé par Schleicher, Haeckel argue ainsi sur l'évidente nécessité d'établir une phylogénie dans le règne animal, puisque les langues s'organisent elles-mêmes de façon phylogénétique:
- 14 « Suivez, à l'aide de cet arbre généalogique (fig. 2), le développement des divers rameaux linguistiques sortis des communes racines de la langue indo-germanique primitive, et vous aurez un tableau extrêmement clair de leur phylogénie [...] Si j'ai quelque peu insisté sur "l'anatomie comparée" et l'histoire du développement des langues, c'est qu'elles éclairent singulièrement la phylogénie des espèces organiques. Vous le voyez, par leur structure et leur évolution, les langues primitives, les langues mères, les langues sœurs et les idiomes répondent très bien aux classes, ordres, genres et espèces du règne animal. Dans les deux cas, la "taxinomie naturelle" est phylogénétique. » (Traduction de la deuxième édition allemande par Letourneau, Reinwald Ed. 1877).

Fig. 2 - Arbre généalogique des langues indo-germaniques, dans sa représentation proposée par Haeckel (1874).

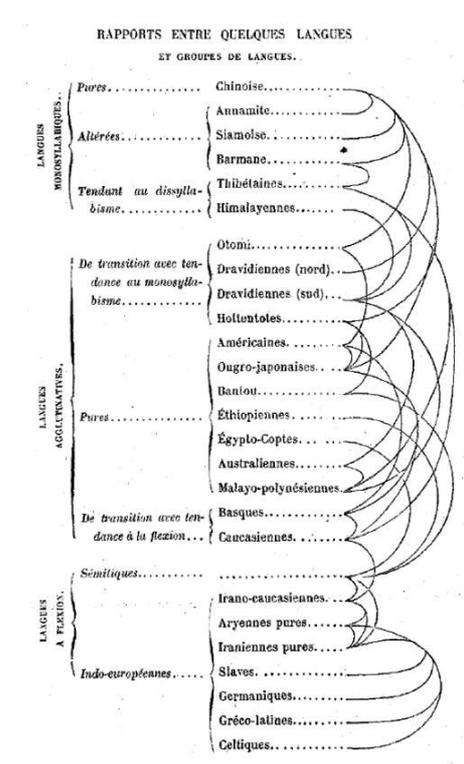
Fig. 2—Genealogical tree of indo-Germanic languages as proposed by Haeckel (1874).



- 15 Les anthropologues de la fin du XIX<sup>e</sup> s. abordèrent également ce problème de l'évolution des langues avec, peut-être, moins de simplisme. Dans son *Introduction à l'étude des Races Humaines* (Quatrefages [de] 1886) de Quatrefages insiste davantage sur la complexité des ressemblances et des rapports qui existent entre les langues que sur leurs relations de parenté. Au point qu'il préfère privilégier une représentation des liens qui unissent les langues entre elles sous forme d'un réseau (fig. 3) plutôt que sous la forme, idéologiquement dominante, d'un arbre dichotomique comme le faisait avant lui Schleicher ou Haeckel.

Fig. 3 - Rapport entre quelques langues selon de Quatrefages (1886) : « Lorsqu'on dispose les langues en tableaux dressés d'après les résultats admis par les maîtres en linguistique, lorsqu'on représente par des lignes les rapports indiqués par eux, on retrouve ici la fusion et l'entrecroisement des caractères, aussi bien que lorsqu'on étudie l'extérieur ou le squelette du corps humain ».

Fig. 3—Relationship between some languages according to de Quatrefages (1886): “When languages are arranged in tables drawn up from results accepted by the greatest authority in linguistics, lines depicting the relationships they described lead to rediscover fusion and intertwinning of characters, as well as when studying physical appearance or skeleton of human body”.



- 16 La source d'une telle prudence, particulièrement chez de Quatrefages, est probablement à rechercher dans la difficulté que rencontraient à cette époque les anthropologues pour élaborer une classification simple et cohérente des races, en tentant de concilier la multiplicité des critères de classement possibles. La situation, mauvaise pour définir les races, n'était pas meilleure pour définir les langues. Sans doute certains anthropologues étaient-ils plus sensibles que les naturalistes à l'importance déterminante des migrations et des échanges, donc aux emprunts en linguistique, qui viennent malencontreusement brouiller les traces d'une histoire naturelle évolutive simplement arborescente.
- 17 Le développement de la génétique, dès la redécouverte des lois de Mendel au début du XX<sup>e</sup> s., marque un tournant dans les relations entre histoire des gènes et histoire des langues. La génétique semble s'être relativement désintéressée de ce type d'analogie, pour déplacer le problème vers des questions fondamentales sur les processus de « transmission » des caractères et donc vers un nouveau paradigme, celui de l'analogie entre transmission des gènes et transmission des faits culturels, incluant les langues. Sewall Wright aborde très brièvement la question, tout comme Dobzhansky qui énumère différentes conceptions sur l'origine du langage dans son livre à succès *L'homme en évolution* (Dobzhansky 1966).

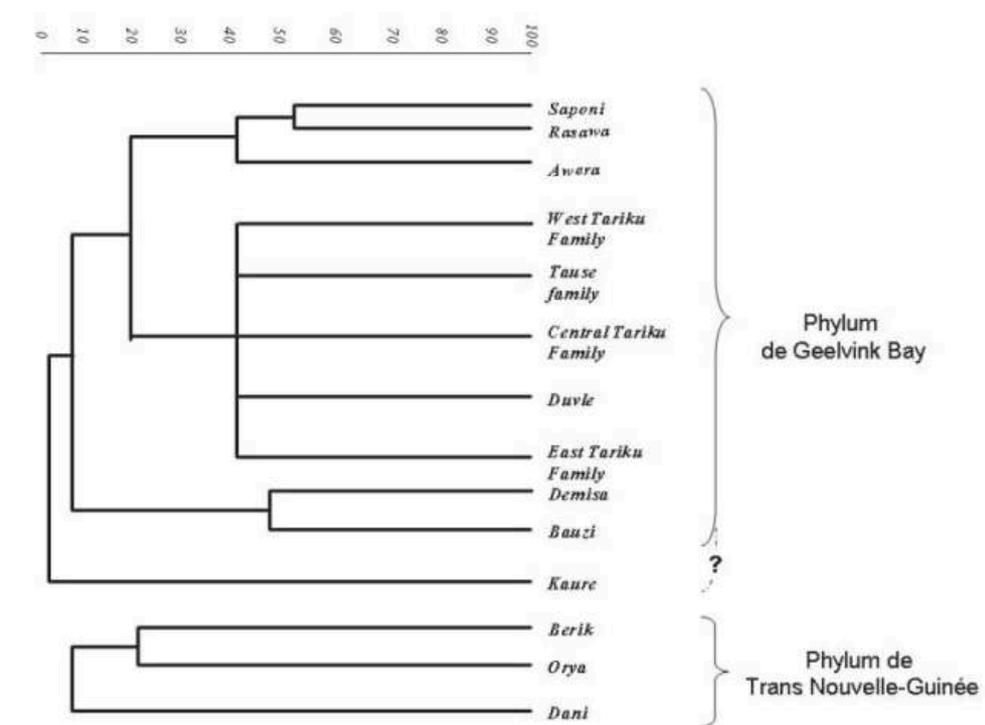
- 18 Cette problématique de la transmission culturelle sera approfondie et formalisée bien plus tard par Cavalli-Sforza et Feldman (Cavalli-Sforza, Feldman 1981) puis par Boyd et Richerson (Boyd, Richerson 1985).

## Les tentatives de la taxonomie numérique

- 19 Il faut attendre les années 1950 pour constater un nouveau rapprochement significatif entre linguistique et génétique, rapprochement consécutif à une convergence méthodologique, lorsque linguistes et biologistes se retrouvèrent sur les mêmes méthodes numériques de classifications. L'application de ces méthodes fut menée parallèlement en biologie systématique (Michener, Sokal 1957 ; Sokal, Sneath 1963 ; Sneath, Sokal 1973) et en linguistique (Ellegard 1959 ; Kroeber 1960 ; Dyen 1962). Leur principe consiste à quantifier la ressemblance entre les objets comparés (espèces, populations ou langues) puis d'en dresser, par un algorithme adéquat, une classification hiérarchique. La question se pose de savoir si de telles méthodes sont capables d'inférer une véritable histoire évolutive, particulièrement au niveau intra-spécifique. Dans leur travail, Michener et Sokal mettent bien en garde sur le fait que la ressemblance ne conduit pas nécessairement à établir des relations phylogénétiques, à moins de justifier la pertinence du choix des caractères, du bien fondé d'un certain nombre d'hypothèses sur les processus d'évolution des objets et des caractères qui les définissent et de la bonne adéquation entre ces hypothèses et la métrique utilisée pour apprécier la ressemblance. De très nombreuses controverses et critiques méthodologiques se sont élevées à cet égard dans les années 1980 (Farris 1972, 1981, 1985, 1986 ; Felsenstein 1984, 1986 ; Fitch 1984). Bien que tous les problèmes ne soient pas totalement résolus, ces méthodes sont toujours utilisées, comme nous le verrons plus loin, dans beaucoup de travaux de génétique des populations humaines.
- 20 Le débat sur la pertinence de ces méthodes numériques a certainement son parallèle dans le domaine linguistique. On en trouve un écho anecdotique dans le travail d'Evrard (Evrard 1966), non linguiste déclaré, qui propose une analyse des affinités linguistiques entre plusieurs dialectes bantous, basées sur la liste de mots de Swadesh. Les affinités entre dialectes, pris deux à deux, sont quantifiées par des coefficients de corrélation dont la matrice est représentée graphiquement. Des groupes compacts de dialectes, repérés dans la matrice, se retrouvent sur des zones géographiques proches, tandis que d'autres éléments de la matrice, correspondant à des corrélations faibles, correspondent à des zones géographiques de transition. Cependant, et comme le souligne prudemment l'auteur : « ... *corrélation ne veut pas dire parenté. Je crains que trop souvent, dans les recherches de ce genre, on ait confondu la corrélation (qui correspond à un état existant au moment des observations) et la parenté ou filiation (qui est un fait historique, dont les traces pourraient disparaître sans que ce fait lui-même cesse d'être vrai). Sans doute peut-on éclairer les recherches d'origine au moyen des études de corrélation. Mais cela suppose une interprétation linguistique des faits sur lesquels l'analyse quantitative a attiré l'attention.* »
- 21 Tout un courant de la linguistique, issu des tentatives de glottochronologie proposées par Swadesh, n'hésite cependant pas à s'affranchir des réserves nombreuses que suscitent ces méthodes quantitatives et à reconstruire des arbres fondés sur une

quantification du degré de cognats partagés entre langues. Un exemple nous est donné par Clouse (Clouse 1996) à propos de langues de Nouvelle-Guinée (*fig. 4*). En dialectologie également, cette approche est courante, si l'on veut bien admettre qu'elle n'est pas seulement intéressée à détecter des variations spatiales des langues, mais aussi à inférer l'histoire de leur genèse. Dans cette optique, Goebel a élaboré toute une stratégie méthodologique d'abord pour construire des matrices de similarité entre formes dialectales et en déduire ensuite des représentations arborescentes. Ses applications concernent l'Italie du Nord (Goebel 1981) ou les dialectes d'Oïl (Goebel 1983, 2004) par exemple. D'autres méthodes se fondent sur les théories de l'information pour calculer un autre type de distance, dite de Levenshtein. Elle mesure le coût du passage d'une forme d'un mot à un autre, par insertion, délétion et/ou remplacement de consonnes ou voyelles. Les matrices de distances servent ensuite à différentes représentations classiques en analyse de données (analyses multivariées), mais également à des représentations sous forme d'arbre, généralement par les méthodes taxonomiques classiques évoquées plus haut. Un exemple est proposé, à la suite des travaux de Kessler sur le Gaélic (Kessler 1995), par Nerbonne *et al.* (1999) à propos des dialectes parlés aux Pays-Bas.

Fig. 4 - Exemple de représentation d'arbre à partir d'une matrice de ressemblance globale (proportion de cognats partagés, langues de Nouvelle-Guinée) (Clouse 1996). Des langues ayant entre 75 % et 40 % de cognats partagés sont considérées comme appartenant à la même famille. L'appartenance à des phylums différents se fait à moins de 5 % de ressemblance entre les langues. Fig. 4—Example of a tree representation of New Guinea languages, drawn from a resemblance matrix based on the pairwise proportion of shared cognates (Clouse 1996). Languages sharing between 75% and 40% of cognates are considered to belong to the same family. When the resemblance between two languages is less than 5%, they are said to belong to two different phyla.

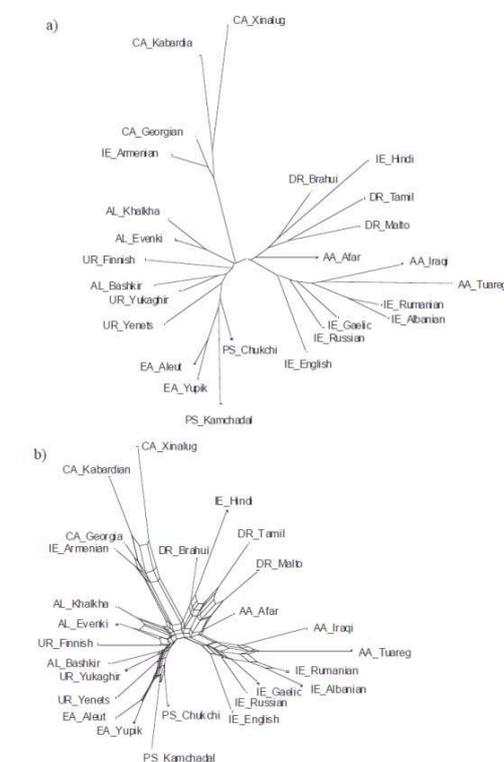


- 22 D'autres approches plus récentes se proposent de mieux extirper l'information contenue dans les matrices de distances entre langues, en se libérant de la contrainte de la représentation stricte en forme d'arbre, et en acceptant une représentation moins contrainte en forme de réseau (Bryant, Moulton 2004). Les « boucles » dans de tels

réseaux rendent alors compte du degré d'incertitude dans les parentés, mais sans que l'on puisse décrypter, pour autant, les raisons de ces incertitudes. On peut en trouver des exemples d'application dans la reconstruction des relations entre langues germaniques, incluant l'anglais et le créole Sarnan (Bryant *et al.* 2005), entre dialectes chinois (Ben Hamed 2005) ou entre quelques langues du monde (*fig. 5*).

**Fig. 5 - Représentation de la matrice des distances entre langues Afro-asiatiques (AA), Indo-européennes (IE), Caucasiennes (CA), Dravidiennes (DR), Altaïques (AL), Paléo-sibériennes (PS), Uraliques (UR) et Eskimo-Aéoutiennes (EA), calculées à partir de 274 données morphologiques et phonologiques, codées en présence/absence et extraites de Ruhlen (1976). Les figures 5a et 5b sont deux représentations par la méthode du Neighbor Joining (Saitou, Nei 1987)) et celle du Neighbor Net (Bryant, Moulton 2004), respectivement. On notera les concordances entre les deux présentations, les regroupements des langues de même famille, le manque de résolution des familles entre elles et les exceptions géographiques de l'Arménien et de l'Hindi.**

**Fig. 5—Representation of the pairwise distance matrix between Afro-Asiatic (AA), Indo-European (IE), Caucasian (CA), Dravidian (DR), Altaic (AL), Paleo-Siberian (PS), Uralic (UR), and Eskimo-Aleut (EA), estimated from 274 morphological and phonological data coded as presence or absence. The data are from Ruhlen (1976). Figure 5a and b are two different representations, by Neighbour Joining (Saitou, Nei 1987) and by Neighbour Net (Bryant, Moulton 2004), respectively. It is worth noting the concordance between these two diagrams, i.e. the clustering of languages that belong to the same family, the lack of discrimination between languages within the families, and the geographical exception of Armenian and Hindi.**



## Les approches cladistiques

- 23 L'introduction des méthodes cladistiques dans le domaine de la systématique a complètement bouleversé la façon d'appréhender la diversité biologique, par un retour aux sources du darwinisme. Le problème n'est plus de quantifier un degré de ressemblance globale entre les espèces (ou entre langues) pour en tirer une représentation sous forme d'arbre, comme on vient de le voir avec les méthodes de taxonomie numérique, mais bien d'analyser la façon dont les caractères se distribuent

entre les différentes espèces (ou langues), en cherchant à identifier les formes ancestrales et les formes dérivées, puis en fondant l'apparement entre espèces sur le partage des formes dérivées des caractères. Les travaux de l'entomologiste Hennig (Hennig 1950) furent déterminants dans la formalisation et la diffusion des concepts de la cladistique. Comme pour les méthodes de taxonomie numérique, le résultat final reste un arbre. Mais l'information qu'il apporte est cependant radicalement différente, puisque il est construit à partir d'une réflexion détaillée sur les caractères – morphologie, ADN ou protéines – et sur les modalités de leur évolution. La méthode exige d'abord une définition claire des caractères, des différentes formes qu'ils peuvent prendre, des modalités de leur transformation et du « coût évolutif » de ces transformations. L'utilisation d'algorithmes de parcimonie permet ensuite d'obtenir un arbre dont les relations entre espèces minimisent le nombre et le coût de ces transformations et procède, simultanément, à leur localisation optimale sur l'arbre, retraçant ainsi les formes des caractères issus d'un ancêtre commun et les distinguant de celles qui sont apparues indépendamment par convergence, transferts horizontaux, réversion, ce que l'on regroupe sous le terme d'homoplasie. Dans l'approche cladistique, c'est la construction de l'arbre par application de la parcimonie qui permet, *a posteriori*, de valider les hypothèses d'homologie sur les caractères.

- 24 Ces méthodes fournissent à l'évidence un outil conceptuel indispensable pour comprendre l'évolution des espèces et de leurs caractéristiques morphologiques ou génétiques. Mais, plus généralement, elles se trouvent adaptées chaque fois qu'il s'agit de rendre compte de l'évolution d'entités définies par des attributs qui se modifient au cours du temps. C'est le cas, par exemple, des manuscrits dont les éditions successives dérivent les unes des autres par copies manuelles successives (Cameron 1987). Ce pourrait être, également, le cas de l'évolution des langues comme on va le voir.

## De l'usage des concepts phylogénétiques en linguistique

### Des analogies terminologiques et conceptuelles

- 25 Le renouveau cladistique de la phylogénie a également joué un véritable rôle de catalyseur dans le rapprochement entre linguistique et biologie. Cela se fit d'abord sous la forme de réunions initiées par des linguistes et auxquelles des généticiens des populations furent conviés, tel le symposium organisé par Hoenigswald en 1982 et 1983, et dont les actes furent publiés en 1987 (Hoenigswald, Wiener 1987), sous le titre de *Biological Metaphor and Cladistic Classification*. Deux autres séminaires suivirent, sur des thématiques analogues : *Cultural transmission and change*, Stanford, 1983 ; et *International Meeting on Language Change and Biological Evolution*, Turin, 1988. On peut dès lors s'interroger sur le produit de ces interactions, en termes de co-optation des concepts et méthodes phylogénétiques par les linguistes pour aborder leurs propres problématiques évolutives, et ce d'autant plus que ces méthodes ont connu au cours des dernières décennies de remarquables développements.
- 26 Dans la préface de son livre, Hoenigswald (1987 : page xi) insiste sur le fait que la méthode comparative de la linguistique historique a été cladistique près d'un siècle avant que ce concept ne soit formulé par Hennig en biologie. Il soutient que le

phylogénéticien qui se pencherait sur la méthode comparative serait surpris de l'étendue des analogies qu'il pourrait dresser avec ses méthodologies propres. C'est un linguiste qui dressera ces analogies. Roger Lass (Lass 1997) considère que celles-ci ne tiennent qu'à partir du moment où la démarche hypothético-déductive se met en place avec 1) la notion de régularité et 2) l'utilisation de la notion d'innovation partagée comme principe de reconstruction. Ces deux points de méthodes, sur lesquels nous reviendrons plus tard, furent introduits en 1878 par Osthoff et Brugmann (Osthoff, Brugmann 1878), dans ce qui est considéré être l'acte fondateur de la méthode comparative scientifique.

- 27 En effet, Lass fait remarquer que l'analogie ne tient pas pour la méthode comparative à ses débuts, qui fondait alors la notion d'affiliation génétique des langues sur leur ressemblance globale ou même sur des considérations culturelles propres aux communautés d'individus plutôt qu'à leurs langues. Mais les ressemblances entre langues ne sont qu'en partie le produit de leur apparentement. En effet, elles peuvent également être le fait du hasard du changement linguistique, la marque de contraintes universelles ou de convergences ou encore le fait d'emprunts au cours des contacts entre langues.
- 28 C'est la notion de régularité qui permet à la méthode comparative de distinguer entre les ressemblances liées à l'ascendance commune (homologies) et les ressemblances d'origine non génétique (homoplasies). Ce principe, autrement appelé *hypothèse néogrammairienne*, identifie le changement linguistique comme un processus régulier, affectant de façon systématique l'ensemble du lexique. Inversement, les phénomènes homoplasiques tels que le hasard ou l'emprunt n'auront d'effet que localement. Par conséquent, une correspondance phonologique entre deux langues n'a valeur d'homologie que si sa régularité à travers le lexique peut être démontrée. Si elle n'apparaît pas de façon systématique, l'hypothèse d'homologie se trouve falsifiée, et la ressemblance observée ne peut servir à diagnostiquer une affiliation génétique entre les langues étudiées.
- 29 Dans son parallélisme méthodologique, Lass (1997) utilise abondamment la terminologie cladistique, mais sans définir pour autant les limites de l'analogie. S'il insiste sur la difficulté à poser les hypothèses d'homologie et à polariser les transformations pour les données linguistiques, il ne soulève pas les problèmes de l'algorithmique de reconstruction de l'arbre, ni les questions de la validation de l'arbre inféré et de la possibilité de contradiction ou d'incongruence entre données multiples (lexique, morphologie, syntaxe); autant de questions qui sont au cœur des développements récents de la phylogénie moderne.

## Des différences opératoires

- 30 « C'est pourtant en ces termes que les méthodes cladistique et comparative diffèrent. Si elles procèdent toutes deux, par le test d'une hypothèse d'homologie, en pratique, elles diffèrent dans leur objectif. » En effet, suivant la proposition néogrammairienne, une fois les homologies diagnostiquées, il s'agit d'identifier les innovations, et de proposer des hypothèses d'apparentement entre langues sur la base du partage d'innovations. L'identification des innovations peut s'appuyer sur des principes de naturalité des processus phonologiques ou sur la fréquence d'occurrence des états de caractère dans le groupe étudié, l'état le plus fréquent étant alors supposé ancestral. En pratique,

toutefois, cette étape est problématique, et les régularités mises en évidence servent à proposer des hypothèses d'état ancestral (proto-langue) plus qu'à définir des hypothèses d'apparements. Alors que la méthode cladistique passe nécessairement par la reconstruction d'un arbre pour accomplir son projet phylogénétique, la méthode comparative n'y a pas nécessairement recours. Dans le premier cas, l'arbre est estimé et sert à la fois à définir les hypothèses d'apparements et à tester les hypothèses d'homologie. Dans le cas de la méthode comparative, l'arbre est optionnel, et lorsque celui-ci est proposé, le choix d'une topologie particulière au détriment des autres (nombreuses) topologies possibles n'est pas argumenté.

- 31 Traditionnellement, la méthode comparative n'est appliquée qu'à l'intérieur d'une même famille et non entre familles linguistiques. En effet, la majorité des linguistes comparatistes s'accordent à penser qu'au delà d'une certaine profondeur temporelle, le signal phylogénétique est irrécupérable. Notamment, l'utilisation de lexiques reconstruits comme point de départ d'une comparaison entre familles entraînerait une incertitude dans la reconstruction des proto-formes trop grande pour attribuer une quelconque fiabilité aux apparements qui en seraient inférés. Cette profondeur temporelle critique est estimée à quelque 10 000 ans dans les cas les plus favorables, mais plus vraisemblablement aux alentours de 6000 ans (Trask 1996), alors même que l'émergence du langage dans l'espèce humaine est suggérée aux alentours de 100 000 ans.
- 32 Une minorité de linguistes s'est toutefois aventurée sur ce terrain hasardeux. En effet, la question de l'extension de la comparaison au niveau suprafamilial s'inscrit dans un débat plus large sur la validité d'approches alternatives au comparatisme classique, et plus spécifiquement de la méthode de comparaison multilatérale - *long range classifications* ou *mass comparison* - proposée par Greenberg (Greenberg 1987). Cette méthode se propose de réduire l'ampleur du travail de comparatisme et de dépasser la limitation qui en découle sur la taille de l'échantillon linguistique prospecté. Il s'agit de constituer, pour l'ensemble des langues considérées, un échantillon de quelques centaines à quelques milliers d'items lexicaux et de les aligner comme on le ferait pour des séquences moléculaires. Plus les alignements sont ressemblants, plus les langues correspondantes seront génétiquement proches, d'où des hypothèses d'apparement entre les langues étudiées.
- 33 Ruhlen, connu pour sa théorie de la monogénèse des langues actuelles (Ruhlen 1994), considère que « *la comparaison des familles de langues n'est pas vraiment différente de la comparaison des langues elles-mêmes* », même s'il reconnaît des difficultés méthodologiques particulières à cette entreprise. Selon lui, le problème de la comparaison entre familles est plutôt lié à la qualité inégale des registres étymologiques reconstruits pour les différentes familles. Certains sont bien fournis, comme pour les familles indo-européenne et ouralique, et permettent la reconstruction des proto-formes, alors que cela n'est pas possible pour d'autres familles. Il suggère alors de considérer les racines lorsque celles-ci existent, sinon de « *choisir celle des formes modernes qui semble la plus représentative des sons et du sens de la racine* ». En utilisant cette approche, Ruhlen (1994) a reconstruit 27 racines pour la langue-mère universelle putative dont il défend l'existence.

## Une adhésion limitée de la part des linguistes

- 34 Ce type d'approche est considéré comme rétrograde par Lass (1997), qui insiste sur l'importance de la détermination des correspondances systématiques pour que les apparentements et les reconstructions proposées aient un sens. Ringe (1999) considère toute approche de l'évolution des langues qui exclurait cette étape comme non scientifique. Pour lui, l'approche de Ruhlen est uniquement typologique, et qui plus est, saturée par les tendances universelles. Toutefois, le phylogénéticien qui se pencherait sur les travaux de Ruhlen (1994) reconnaîtrait sa méthodologie comme cladistique, même si l'auteur ne s'y réfère pas explicitement et même s'il n'utilise aucun des algorithmes correspondants pour produire ses inférences. De même, la méthode de comparaison multilatérale, en alignant les items lexicaux de même valeur sémantique suggère leur traitement comme des séquences moléculaires : chaque segment du mot occupe un locus pouvant avoir un certain nombre d'états possibles et a valeur de caractère. Alors que la communauté comparatiste rejette en bloc les conclusions de Ruhlen, rejetées également par les analyses probabilistes de ses données (Boë *et al.* 1994, 2003), celles-ci suscitent l'intérêt des généticiens des populations. En effet, outre l'évocation d'une méthodologie très semblable à celles pratiquées en phylogénie conduisant à une classification facile à manipuler, les conclusions de Ruhlen coïncident avec l'hypothèse d'un goulot d'étranglement démographique à l'origine des populations humaines actuelles.
- 35 Étrangement, malgré le parallélisme flagrant entre les méthodologies développées en linguistique et en biologie pour adresser les questions d'évolution, les échanges entre linguistes et biologistes sur les points de méthodes – convergents ou divergents – ont été très rares et sont très récents (Forster, Renfrew 2006). Le pont tendu par Lass (1997) entre linguistique historique et cladistique est pour le moment la seule tentative explicite élaborée par un linguiste et elle n'a pas débouché sur des applications concrètes de la part de linguistes comparatistes. Ruhlen (1994), bien que pionnier par son discours cladiste, n'applique pas rigoureusement, ni jusqu'au bout, la méthode cladistique à ses données.
- 36 Ringe *et al.* (Ringe 1999 ; Ringe *et al.* 2002) ont bien tenté d'aborder la question d'estimation de l'arbre dans le cas de la famille indo-européenne, mais ils utilisent pour cela une approche dite de compatibilité, qui consiste à identifier le plus grand ensemble de caractères compatibles hiérarchiquement, et d'estimer les arbres correspondants. Une telle approche évacue d'emblée le problème de l'homoplasie puisqu'elle ignore les caractères incompatibles et, par conséquent, elle ne permet pas de la quantifier ni de la caractériser. Les auteurs ont cependant reconnu que si les majorités de leurs caractères étudiés étaient compatibles avec l'arbre résultant de leur approche, certains autres ne l'étaient pas. De là, ils ont étendu leur approche à la production de réseaux, autorisant ainsi la présence de réticulations localement dans l'arbre, pour rendre compte des caractères incompatibles avec le modèle arboré parfait, et ce faisant, les emprunts entre langues (Nakhleh *et al.* 2005). Pour ce faire, les caractères non compatibles sont ajoutés à l'arbre précédemment obtenu par une méthode de compatibilité, de sorte à minimiser le nombre de réticulations à constituer pour en rendre compte. Le postulat est donc fait que l'arbre qui approxime le mieux la phylogénie vraie est nécessairement celui avec lequel le plus grand nombre de caractères est compatible. Les caractères incompatibles n'interviendront pas dans sa construction mais seront placés à

*posteriori* pour lier entre elles les branches où des emprunts sont supposés avoir eu lieu.

- 37 D'autres approches, plus clairement conformes à l'esprit de la cladistique, se polarisent essentiellement et en premier lieu, sur la meilleure façon d'exprimer, sous forme de codage, les hypothèses évolutives et les postulats formulés par les linguistes, avant tout traitement algorithmique. Les exemples concernent pour l'instant les changements vocaliques au niveau eurasiatique (Ben Hamed *et al.* 2005) ou au niveau dialectal (Gaillard-Corvaglia *et al.* 2007).
- 38 Enfin, les développements probabilistes récents, exigés pour bien rendre compte de l'évolution des séquences d'ADN ou de protéines, ont également connu un succès certain à l'occasion de leur transfert à des données linguistiques. L'intérêt de ces méthodes est de pouvoir complexifier les modèles, mais reste généralement relativement réduit en ce qui concerne l'aspect linguistique. Les données consistent généralement en des listes lexicales (liste de Swadesh ou Dyen) codées en présence/absence et dont les probabilités de changements sont estimées par optimisation en même temps que l'arbre, par maximum de vraisemblance ou par des méthodes bayésiennes. L'avantage de ces méthodes, une fois acceptée la traduction « probabiliste » des modifications linguistiques, est de permettre d'effectuer des tests statistiques à la fois sur la robustesse de l'arbre, les degrés de divergences des langues et les dates aux nœuds de divergence. Un exemple récent porte sur l'Indo-européen (Gray, Atkinson 2003 ; Atkinson *et al.* 2005).

## Phylogénies et réseaux linguistiques et tests d'hypothèses de peuplement

- 39 Récemment, des phylogénies de langues ont également été utilisées pour tester des hypothèses démiques, culturelles ou historiques. Cette perspective se nourrit de la pratique, déjà étreinée en écologie ou en psychologie évolutive, d'utiliser des phylogénies à des fins de tests d'hypothèses biogéographiques ou comportementales. Ce type d'approche est novateur par rapport aux approches précitées dans le sens où elles s'appliquent à des données linguistiques : les arbres de langues sont d'abord reconstruits par des méthodes phylogénétiques, cladistiques ou autres, puis utilisés pour tester des hypothèses ou des scénarios évolutifs non linguistiques.
- 40 Ainsi, Gray et Jordan (2000) ont voulu tester les scénarios alternatifs envisagés pour le peuplement du Pacifique sur la base d'une phylogénie linguistique de 77 langues austronésiennes reconstruite par parcimonie à partir de listes lexicales. Le premier scénario de peuplement proposait une expansion d'île en île, d'ouest en est à partir de Taïwan, impliquant une différenciation linguistique et génétique des populations du fait de l'isolement par la distance imposé par la géographie. Le scénario alternatif suggérait, au contraire, une origine mélanésienne et une dispersion des populations sans isolement relatif du fait d'un treillis socio-politico-économique les maintenant en contact, impliquant des contacts entre langues tout au long du peuplement de cette région. Selon le scénario envisagé, la forme de l'arbre des langues est supposée différer. En l'occurrence, celle-ci montre une séquence géographique d'ouest en est et une origine taïwanaise, en accord avec le premier scénario présenté, mais en contradiction avec l'arbre attendu selon le second scénario qui devrait n'avoir aucune structure géographique puisque les relations entre langues sont supposées être structurées en

réseau et non en arbre. Par la suite, Greenhill (2002) et Greenhill et Gray (2005) ont raffiné l'utilisation de ces phylogénies linguistiques pour tester statistiquement des hypothèses sur le peuplement du Pacifique, plus complexes ou intermédiaires entre les deux scénarios précédents et tenant compte, notamment, de la distribution des caractères sur les arbres et de l'incertitude sur les estimations phylogénétiques.

- 41 Holden (2002) a étendu ce type de méthodologie à l'étude de l'évolution de traits culturels, l'objectif étant de préciser dans quelle mesure l'arbre des langues reflète une histoire culturelle plus large des populations locutrices. L'auteur s'est intéressé aux langues bantoues d'Afrique et a testé un scénario de diversification des langues concomitante à l'expansion de l'agriculture. Sous une telle hypothèse, on s'attend à ce que les langues soient reliées entre elles selon un schéma arboré, et que la chronologie relative des nœuds de cet arbre coïncide avec la séquence des artefacts archéologiques laissés par les agriculteurs lors de leur expansion au cours du Néolithique. Son analyse repose sur des listes lexicales pour 75 langues du domaine géolinguistique étudié, et la reconstruction phylogénétique se fait par parcimonie. La topologie consensus obtenue montre une forte composante géographique en accord, dans sa chronologie relative, avec les artefacts archéologiques, corroborant ainsi le scénario de différenciation linguistique du domaine bantou en lien avec l'expansion des agriculteurs au Néolithique.
- 42 Toutefois, ces approches soulèvent un certain nombre de critiques méthodologiques. Ainsi, les arbres obtenus sont rarement très robustes, remettant en cause la pertinence des relations entre langues, ce qui est pourtant le pivot du test d'hypothèse. Par ailleurs, la simplicité des modèles évolutifs utilisés et leur focalisation sur les données lexicales sont également reprochées. Ces deux limitations sont en réalité corrélées. Outre le fait que les données disponibles et accessibles à la compréhension du phylogénéticien, généralement non linguiste, sont exclusivement lexicales, la façon dont elles sont exploitées procède par analogie simple avec la manipulation de données moléculaires peu complexes (comme les quatre états A, T, C, G). Face à des caractères linguistiques infiniment plus polymorphes et pour lesquels aucune théorie évolutive synthétique n'est disponible, le phylogénéticien se retrouve bien moins armé et souvent désemparé, sur un terrain méthodologique qui lui est quelque peu éloigné. Cette difficulté à prendre en compte la complexité linguistique se traduit de façon évidente dans les calculs des distances lexicostatistiques déjà évoqués. En génétique des populations, les distances peuvent rendre compte des différents mécanismes qui influent sur l'évolution du génome d'une population, à savoir la mutation, la migration, la dérive et la sélection. Elles peuvent également intégrer les fluctuations dans la taille des populations et l'hétérogénéité des taux de substitution dans le temps. En revanche, les distances linguistiques sont bien plus réductionnistes, faute de savoir définir des mécanismes généraux quantifiables. Ainsi, le phylogénéticien se trouve enclin à appliquer les spécificités de ses propres modèles moléculaires à une réalité bien différente, celle de l'évolution des langues.
- 43 Ainsi, bien que ces approches aient connu une popularité grandissante au cours des dernières années auprès des biologistes, des anthropologues ou des archéologues, elles ne vont pas sans éveiller, au sein de la communauté linguistique et à de rares exceptions près, des appréciations mitigées, entre intérêt critique et scepticisme, envers des pratiques qui sont, pourtant, fondées sur les analogies entre linguistique et phylogénie. En revanche, les généticiens, plus audacieux ou moins pointilleux, n'ont

pas hésité à intégrer des données linguistiques dans leurs travaux sur la diversité génétique des populations.

## De l'usage de la linguistique en génétique des populations

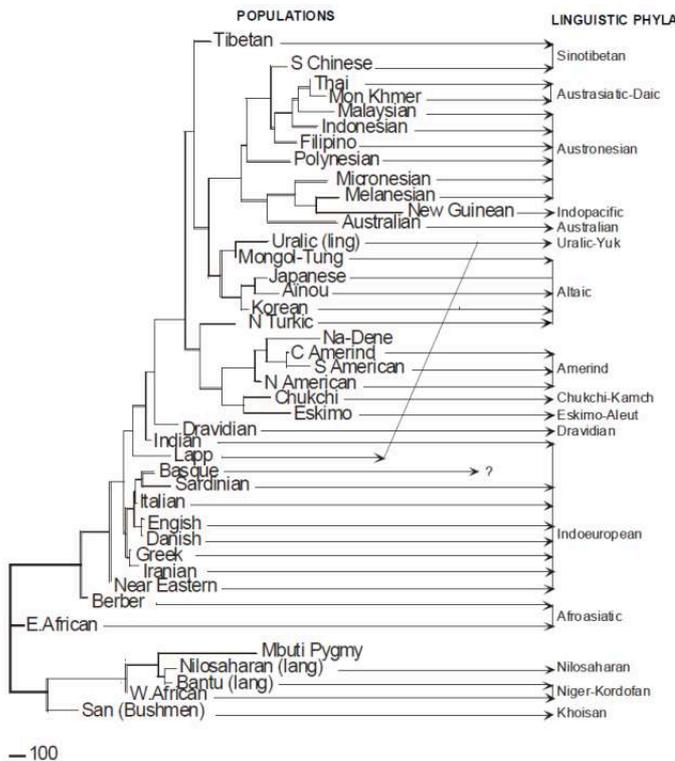
- 44 La localisation géographique commune (Stanford, USA) d'un centre d'excellence en génétique des populations humaines et de l'école linguistique de Greenberg a certainement favorisé l'osmose entre la génétique des populations et la linguistique historique. Il en est ressorti de multiples publications de génétique des populations pour lesquelles la référence linguistique presque exclusive reste la classification proposée par Ruhlen (1987). Ces publications relèvent de trois catégories d'approches. Celles qui retracent une histoire évolutive des populations au travers de leurs gènes et cherchent à établir un parallèle linguistique ; celles qui tentent de quantifier les ressemblances génétiques et linguistiques afin de mesurer leurs corrélations ; enfin, celles qui recherchent, dans l'espace géographique, des coïncidences entre disparités génétiques et discontinuités linguistiques.

### Arbre génétique et langues

- 45 Le travail de Cavalli-Sforza *et al.* (1988) relève de la première catégorie (*fig. 6*). Ces auteurs estiment les distances génétiques entre 42 populations, réparties sur tous les continents, en utilisant la quasi-totalité (120) des polymorphismes génétiques étudiés jusqu'à cette date. L'arbre phylogénétique qu'ils en tirent est obtenu par application d'une méthode de taxonomie numérique classique supposant que toutes les populations évoluent à des vitesses comparables (hypothèse « d'horloge »). L'arbre montre une première dichotomie entre les populations africaines et les autres populations, en accord avec les premières conclusions tirées de l'ADN mitochondrial par Cann *et al.* (1987) et dont est sortie la fameuse hypothèse de « l'Ève Africaine ». Sur cet arbre, une correspondance semble s'établir entre les groupes de populations fondées sur la génétique et leur appartenance linguistique définie à partir de la classification proposée par Ruhlen en 1987. Les auteurs concluent à un « parallélisme considérable entre évolution génétique et linguistique ».

Fig. 6 - Correspondance entre l'arbre phylogénétique des populations humaines, construit sur la base de fréquences géniques, et l'appartenance linguistique dans la classification de Ruhlen (1987). Cette figure se différencie de celle proposée par Cavalli-Sforza et al. (1988) sur plusieurs points. La matrice de distance génétique entre populations est ici la matrice des  $F_{st}$  proposée par Cavalli-Sforza et al. (1994) et non la distance de Nei (Cavalli-Sforza et al. 1988). L'algorithme utilisé est le Neighbor-joining, préservant ainsi la longueur des branches, et non la méthode « Average linkage ». L'origine est placée artificiellement en mettant les populations africaines en situation d'extra-groupe. Les populations européennes (incluant le Basque) ne sont pas « agglomérées ». Tout en conservant la structure de l'arbre, la représentation a été optimisée pour tenter de regrouper les populations selon leur phylum linguistique d'appartenance tel qu'utilisé par Cavalli-Sforza et al. (1988), même lorsque ces populations ne forment pas de groupes monophylétiques.

Fig. 6—Correspondence between the phylogenetic tree of human populations based on gene frequencies, and their linguistic classification (Ruhlen 1987). The figure differs from the one proposed by Cavalli-Sforza et al. 1988) by two points. The genetic distance matrix is the  $F_{st}$  matrix given by Cavalli-Sforza et al. (1994) instead of the Nei's distance (Cavalli-Sforza et al. 1988). The algorithm used to draw the tree is the Neighbour Joining preserving branch lengths, instead of "Average linkage" method. The origin of the tree is artificially placing African population as an outgroup. European populations (including Basques) are not pulled together. The tree representation has been optimised, however taking constant its hierarchical structure, in an attempt to bring as close as possible populations belonging to the same linguistic phylum as in Cavalli-Sforza et al. (1988), even if these populations do not form a cluster.



- 46 Ce travail a essuyé quelques attaques virulentes, en particulier de la part d'O'Grady *et al.* (1989) qui l'estiment pauvrement soutenu empiriquement et méthodologiquement : forte proportion de données manquantes et taille restreinte de certains échantillons, définition contestable des populations, représentation de sept familles linguistiques (sur 16) par une seule population et donc en cohérence avec n'importe quelle structure d'arbre, accord entre populations génétiques et familles de langues pour seulement cinq familles, absence de soutien statistique des conclusions. De plus, O'Grady conteste que l'on puisse considérer la langue comme un caractère évolutif de la population, tandis que Cavalli-Sforza postule que les différenciations linguistiques et génétiques sont soumises aux mêmes facteurs, bien qu'elles ne soient pas identiques dans leur processus. Suite à une suggestion d'O'Grady *et al.* (1989), Cavalli-Sforza *et al.* (1992)

démontrent par un test de co-évolution, la présence d'« association non aléatoire entre familles linguistiques et histoire génétique humaine ».

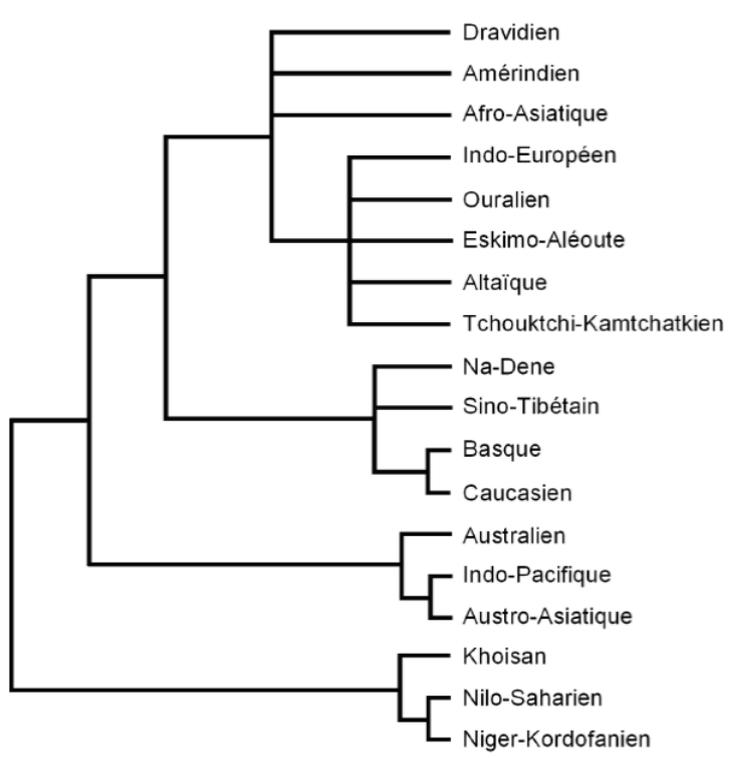
- 47 De leur côté, Bateman *et al.* (1990) soulignent l'incohérence d'une comparaison entre histoire, celle de l'arbre génétique des populations, et classification purement typologique des langues. Ils insistent sur la nécessité d'effectuer indépendamment une reconstruction « historique » des relations de parenté entre langues, par une approche cladistique. Pour eux, les conclusions de Cavalli-Sforza sur le parallélisme entre évolution des langues et des gènes ne seront véritablement probantes que lorsqu'une co-évolution sera effectivement démontrée. Quelques coïncidences entre histoire évolutive des gènes et typologie des langues ne constituent pas une preuve de congruence.

### Distance génétique, distance linguistique

- 48 La finalité de cette deuxième catégorie de travaux est de vérifier si les différences génétiques entre populations s'accroissent en même temps que leur différence linguistique ou que la distance géographique qui les sépare. Le problème est donc, en premier lieu, de définir des distances linguistiques. Ce point est généralement un sujet de controverse pour les linguistes mais ne préoccupe pas davantage les généticiens, habitués qu'ils sont à manipuler sans complexes des distances. En second lieu, il s'agit d'estimer la corrélation entre les distances génétique, linguistique et géographique et d'en discuter la signification.
- 49 En 1976, Chakraborty (1976) utilisait déjà une telle approche, à partir des classifications hiérarchiques des langues proposées par Greenberg (1960), pour expliquer la diversité génétique des Indiens des Hauts Plateaux andins. La corrélation entre distances génétiques et linguistiques s'avérait quasi nulle ( $r = 0.03$ ). Dans un contexte géographiquement plus large, Chen *et al.* (1995) proposent, une estimation quantitative de la « distance linguistique » entre populations, sur une échelle arbitraire entre 1 ou 2 pour des langues similaires (par exemple langues germaniques), et 60 pour des langues appartenant à deux « phyla » linguistiques totalement différents, la définition des phyla et des familles linguistiques étant toujours tirée de Ruhlen. Les auteurs reconnaissent qu'il s'agit d'une distance « subjective », mais qui répondrait, selon eux, à un certain réalisme. Leur travail, qui procède d'une approche comparable à celle de Cavalli-Sforza *et al.* (1988) mais incluant plus de populations (130), plus de langues (117) et moins de marqueurs (33), conduit à proposer une phylogénie des langues déduite, comme le suggérait Darwin, de la phylogénie des populations (*fig. 7*). Enfin, la corrélation estimée sur ces données entre distances génétique et linguistique reste relativement faible ( $r = 0.22$ ), bien que significative, une fois tenu compte de la distance géographique. Autrement dit, moins de 5 % de la diversité linguistique s'expliquent par des différences génétiques.

Fig. 7 - Représentation des appartenances linguistiques de populations dont l'arbre phylogénétique est construit à partir de fréquences géniques par une méthode de taxonomie numérique (adapté de Chen et al. 1995).

Fig. 7—Description of the linguistic families of populations for which the tree genealogy is based on gene frequencies distances and obtained by a numerical taxonomy method (from Chen et al. 1995).



- 50 En 1988, Sokal (Sokal 1988) appliqua également cette méthode aux données génétiques collationnées en Europe et démontra que la distance génétique entre populations s'explique à la fois par la distance géographique et par la distance linguistique. Cette dernière, toujours fondée sur l'atlas linguistique de Ruhlen, est considérée comme nulle pour deux langues appartenant à une même famille, égale à 1 pour deux langues appartenant à un même phylum mais à deux familles différentes, et à 2 pour deux langues appartenant à deux phyla distincts. D'autres essais de quantification de la distance, sur une échelle de 1 à 60, toujours sur la base des données de Ruhlen, ont été proposés par Poloni *et al.* (1997). Cette distance est mise en corrélation avec des distances génétiques fondées sur le polymorphisme du chromosome Y, de l'ADN<sub>mt</sub> et des données RFLP. La corrélation entre distances génétique et linguistique, à distance géographique constante, reste modeste ( $r = 0.32$ ). Dans un travail sur les populations d'Asie et du Pacifique, Lum *et al.* (1998) proposent une distance linguistique sur une échelle de 1 à 6 qui, cette fois, n'est pas fondée sur la classification de Ruhlen, mais sur la proportion d'innovations ou de cognats partagés.
- 51 Très récemment, Wood *et al.* (2005) ont appliqué la même méthodologie au continent africain. La corrélation entre distances linguistique et génétique s'avère plus forte quand cette dernière est estimée à partir du polymorphisme du chromosome Y, transmis par les hommes ( $r = 0.33$ ), que lorsqu'elle est estimée à partir de l'ADN mitochondrial, transmis par les femmes ( $r = 0.17$ ). Le rôle des Bantu serait déterminant dans ce contraste de corrélations dans la mesure où leur rapport s'inverse lorsqu'ils ne

sont pas pris en compte dans l'analyse. Sans doute un effet de migrations différentielles des hommes et des femmes lors de l'expansion Bantu.

52 Plus récemment encore, et pour répondre aux analyses précédentes jugées partielles, Belle et Barbujani (2007) ont proposé d'affiner les résultats en se basant sur une meilleure couverture mondiale de la diversité génétique, en utilisant des données sur les microsatellites dont les taux élevés de mutation les rapprocheraient, au dire des auteurs, des taux de changement observés en linguistique. Par ailleurs, le nombre conséquent (377) de microsatellites pris en considération, dont la fréquence est relevée sur 50 populations mondiales, donne plus de poids aux conclusions. Cependant, les distances linguistiques entre populations restent frustes : distance de 1 entre langues, de 2 entre familles de langues, de 3 entre branches et de 4 entre phyla, fondées essentiellement sur l'Atlas linguistique de Ruhlen, complétées, lorsque nécessaire, par les classifications ethnologiques du site « *The Ethnologue* » (Gordon 2005). La conclusion de cette analyse conforte la faible relation, au niveau mondial, entre diversité génétique et diversité linguistique, une fois prises en compte les distances géographiques.

Pas plus de 3 % de la diversité génétique s'expliquerait par la diversité linguistique.

53 Ces différentes tentatives de corréler distance génétique et distance linguistique conduisent à des résultats éminemment variables, selon les populations considérées, leur dispersion géographique, le matériel génétique étudié et la métrique utilisée pour définir la ressemblance linguistique. Tirer des conclusions de telles études reste donc fortement conjoncturel. De plus, la faiblesse rédhitoire de ces approches tient au fait que la présence d'une corrélation entre langues et gènes ne constitue aucunement la preuve d'une histoire évolutive convergente. C'est à la fois une question de statistique, – une corrélation n'établit pas nécessairement un lien de causalité direct –, et de méthode, – une corrélation n'est pas une congruence au sens phylogénétique du terme

## Frontières génétique et linguistique

54 Devant les résultats relativement limités de toutes ces confrontations entre données génétiques et linguistiques, de nouvelles approches se sont développées, se focalisant sur la recherche de discontinuités spatiales dans la variation génétique des populations. Il s'agit de vérifier si de telles discontinuités n'auraient pas pour origine des discontinuités corrélatives dans la distribution des langues.

55 Barbujani et Sokal (1990) reconnaissent en Europe près de 33 zones géographiques présentant de brusques variations de fréquences géniques, délimitant ainsi des sortes de « frontières » génétiques. Se fondant sur les appartenances linguistiques de l'Atlas linguistique de Ruhlen, ils observent que ces frontières séparent tout autant des populations parlant des langues de familles linguistiques différentes que des populations parlant des langues différentes mais d'une même famille. Comme plus de 70 % de ces frontières coïncident également avec des frontières géographiques, l'explication exclusivement linguistique de la différenciation génétique reste bien précaire.

56 Rosser *et al.* (2000) ne trouvent pas de résultats plus concluants à partir du polymorphisme du chromosome Y de 47 populations européennes et 37 langues définies selon la classification de Ruhlen. Les barrières génétiques sont aussi nombreuses entre populations parlant des langues de différentes familles qu'entre

populations parlant des langues d'une même famille mais de sous-familles différentes, ou même des langues d'une même sous-famille. Ce même travail ne retrouve pas de corrélation entre distance génétique et distance linguistique à distance géographique constante ( $r = 0.09$ ). Leur distance linguistique, adaptée de Dyen (Dyen *et al.* 1992) est basée sur la méthode lexicos-tatistique de Swadesh (1952).

- 57 Cette absence de corrélation, au niveau européen, est en contradiction avec les travaux de Sokal obtenus sur des marqueurs autosomaux. Elle est également en contradiction, probablement pour des raisons d'échelle, avec ceux de Polini *et al.* (1997) toujours sur le chromosome Y mais sur une zone géographique incluant l'Afrique et l'Europe. En restant sur cette même zone géographique et linguistique, mais en se concentrant sélectivement sur le seul système génétique Rhésus, Dupanloup *et al.* (2000) montrent quel'effet « linguistique » sur la barrière génétique entre populations afro-asiatiques et indo-européennes ne serait que secondaire à un effet géographique.
- 58 Quittant l'Europe pour l'Amérique du Sud, Hunley et Long (2005) constatent également qu'à leur échelle continentale, il n'y a pas de recouvrement entre la classification linguistique de Greenberg et la diversité génétique pour l'ADN mitochondrial.
- 59 Le travail plus récent de Belle et Barbujani (2007), déjà évoqué plus haut, applique aux populations mondiales la même méthodologie. Il s'agit toujours de repérer de possibles frontières génétiques entre leurs 50 populations, par des méthodes de triangulation (Manni *et al.* 2004) ; ensuite de vérifier quel degré de divergence linguistique, selon la classification de Ruhlen, s'observe de part et d'autre de ces frontières. Il s'avère que les frontières génétiques séparent statistiquement ( $P < 0.05$ ) davantage de langues classées dans des phyla distincts que de langues appartenant au même phylum. Cependant, cette conclusion reste ténue dans la mesure où n'est pas pris en compte le fait que ces frontières génétiques n'ont pas toutes la même « intensité » ni la même validité statistique. D'ailleurs, pour augmenter la validité statistique de leur conclusion, les auteurs n'hésitent pas à s'écarter de la classification proposée par Ruhlen pour l'Amérique du Sud (un seul phylum) pour adopter celle de « *The Ethnologue* » (Gordon 2005) qui retient trois phyla en Amérique du Sud au lieu d'un seul. Cela permet de rajouter trois barrières correspondant à ces trois phyla et, par ce procédé *ad hoc*, de relever la signification du test statistique à  $P < 0.01$ ... Ce jeu avec les classifications conforte la prudence des auteurs qui soulignent que la classification des langues à l'échelle du globe étant encore controversée, il est préférable d'attendre qu'il se dégage un consensus à leur propos, avant de se prononcer sur ce problème de corrélations entre gènes et langues...

## Conclusion

- 60 Au terme de cette longue histoire, commencée dès le XIX<sup>e</sup> s. et qui relate les nombreuses interférences entre approches linguistiques et naturalistes de l'évolution de l'Homme, l'impression domine d'une histoire inachevée, toujours en train de s'écrire mais toujours semée de malentendus. Le lien fondamental entre ces deux disciplines vient de ce qu'elles partagent un but ultime, celui de rendre compte de l'histoire de l'humanité : une origine unique suivie de diversifications biologique et linguistique concomitantes par isolation des groupes les uns des autres lors de leur essaimage sur l'immensité des terres. Ce scénario évolutif se traduit naturellement par une représentation en forme d'arbre généalogique, telle que l'ont préconisée à la fois les

linguistes et les naturalistes du XIX<sup>e</sup> s. Ce type de représentation provoqua, dans les deux disciplines, le développement et la mise en commun, plus ou moins explicite, de tout un ensemble de méthodologies de reconstruction d'arbres, essentiellement les méthodes numériques de classification et les méthodes cladistiques (Forster, Renfrew 2006). Cependant, cette apparente homologie entre les mécanismes évolutifs des langues et des gènes et le partage des mêmes méthodologies pour leur exploration ne doivent pourtant pas camoufler les grandes disparités entre linguistique et génétique. Elles sont d'ailleurs fréquemment discutées par les linguistes (Laks 2002, 2006 ; Forster, Renfrew 2006 ; Szulmajster-Celnikier 2006). Par exemple, si les emprunts entre langues peuvent s'apparenter aux échanges de gènes entre populations ou à l'homoplasie des phylogénéticiens, leur importance respective n'est probablement pas comparable. Les vitesses de changement des fréquences géniques ou des taux de mutations de l'ADN n'ont probablement rien de comparable avec celles des changements phono-logiques, morphologiques ou syntaxiques. Dans ces conditions, la recherche de congruence entre histoire des populations et histoire des langues mériterait davantage qu'une simple fréquentation superficielle de deux disciplines telle qu'elle se pratique depuis près de deux siècles. Bien que la construction d'une histoire d'*Homo sapiens* ne puisse s'affranchir d'une approche syncrétique entre ces disciplines, elle réclame, avant tout, une remise en cause définitive d'une métaphore qui n'a pas totalement convaincu. Elle exige ensuite la construction de modèles rendant mieux compte de la complexité respective des évolutions génétique et linguistique, le recueil de données plus élaborées et une confrontation entre histoires reconstruites à différentes échelles de temps et d'espace.

## Remerciements

- 61 Ce travail a bénéficié du financement de l'action OHLL (Origine de l'Homme, du Langage et des Langues du CNRS), projet « Evaluation de la congruence entre les évolutions génétique, morphologique et linguistique d'*Homo sapiens* ».
- 62 **Apomorphie** : *n.f.* qualifie, dans une séquence de transformations évolutives d'un caractère, l'état de ce caractère dérivé d'un état ancestral. On parle alors d'état apomorphe.
- 63 **Caractère** : *n.m.* tout attribut utilisé pour reconnaître, décrire, définir ou différencier des taxons\*.
- 64 **Cladistique (analyse)** : *n.f.* méthode d'analyse des caractères, mise au point par Willi Hennig en 1966, visant à déterminer la séquence évolutive de leur transformations.
- 65 **Convergence** : *n.f.* ressemblance apparue in-dépendamment dans différents taxons\*, par conséquent, non héritée d'un ancêtre commun.
- 66 **Extra-groupe** : *n.m.* groupe que l'on situe, *a priori*, à l'extérieur du groupe au sein duquel on cherche à déterminer les apparentements.
- 67 **Homoplasie** : *n.f.* ressemblance due à des convergences\* ou des reversions.
- 68 **Plésiomorphie** : *n.f.* se dit de l'état ancestral d'un caractère\*.
- 69 **Polarisation** : *n.f.* opération consistant à déterminer, dans une séquence évolutive, l'état plésiomorphe\* et l'état apomorphe d'un caractère.

- 70 **Racine** : *n.m.* point d'origine de l'arbre, permettant son orientation dans le temps. La racine est donc le point de référence pour l'interprétation des caractères : les états de caractères à la racine sont les états plésiomorphes\*, les états qui en diffèrent sont les états apomorphes\*.
- 71 **Synapomorphie** : *n.f.* partage d'apomorphie par deux ou plusieurs taxons\*.
- 72 **Taxon** : *n.m.* groupe d'organismes (ou d'objets) constituant une unité formelle à chacun des niveaux d'une classification hiérarchique.

## BIBLIOGRAPHIE

- Atkinson (Q.D.), Nicholls (G.), Gray (R.D.) 2005, From words to dates: water into wine, mathemagic or phylogenetic inference? *Transactions of the Philological Society* 103: 193-219.
- Barbujani (G.), Sokal (R.R.) 1990, Zones of sharp genetic change in Europe are also linguistic boundaries, *Proceedings of the National Academy of Sciences, USA*, 87: 1816-1819.
- Bateman (R.M.), Goddard (I.), O'Grady (R.), Funk (V.A.), Mooi (R.), Kress (W.J.), Cannell (P.) 1990, Speaking of forked tongues, *Current Anthropology* 31: 1-24.
- Belle (E.M.S.), Barbujani (G.) 2007, Worldwide analysis of multiple microsatellites: language diversity has a detectable influence on DNA diversity, *American Journal of Physical Anthropology* 133: 1137-1146.
- Ben Hamed (M.) 2005, Neighbour-nets portray the Chinese dialect continuum and the linguistic legacy of China's demic history, *Proceedings of the Royal Society B London* 272: 1015-1022.
- Ben Hamed (M.), Darlu (P.), Vallée (N.) 2005, On cladistic reconstruction of linguistic trees through vocalic data, *Journal of Quantitative Linguistics* 12: 79-109.
- Boë (J.-L.), Bessière (P.), Vallée (N.) 2003, When Ruhlen's Mother Tongue theory meets the null hypothesis, in 15th International Congress of Phonetic Sciences, ICPH Ed., Barcelona, p. 2705-2708.
- Boë (J.-L.), Schwartz (J.-L.), Vallée (N.) 1994, The prediction of vowel systems: Perceptual contrast and stability, in E. Keller (ed.), *Fundamentals of Speech Synthesis and Speech Recognition*, Wiley & Sons Ltd, London.
- Boyd (R.), Richerson (P.J.) 1985, *Culture and the Evolutionary Process*, The University of Chicago Press, Chicago.
- Bryant (D.), Moulton (V.) 2004, Neighbor Net: An agglomerative method for the construction of phylogenetic network, *Molecular Biology and Evolution* 21: 255-265.
- Bryant (D.), Filimon (F.), Gray (R.D.) 2005, Untangling our past: languages, trees, splits and networks, in R. Mace, C.J. Holden, S. Shennan (eds), *The Evolution of Cultural Diversity: A Phylogenetic Approaches*, Routledge Cavendish Publishers, p. 67-84.
- Cameron, (H.) 1987, The upside-down cladogram: Problems in manuscript affiliation, in M.H. Hoenigswald,

- F.L. Wiener (eds), *Biological Metaphor and Cladistic Classification: An Interdisciplinary Perspective*, University of Pennsylvania Press, Philadelphia, p. 227-242.
- Cann (R.L.), Stoneking (M.), Wilson (A.C.) 1987, Mitochondrial DNA and human evolution, *Nature* 325, 31-36.
- Cavalli-Sforza (L.L.), Feldman (M.W.) 1981, *Cultural transmission and Evolution: A quantitative approach*, Princeton University Press, Princeton, New Jersey.
- Cavalli-Sforza (L.L.), Menozzi (P.), Piazza (A.) 1994, *The history and Geography of Human Genes*, Princeton University Press, Princeton, New Jersey.
- Cavalli-Sforza (L.L.), Minch (E.), Mountain (J.L.) 1992, Coevolution of genes and languages revisited, *Proceedings of the National Academy of Sciences, USA*, 89: 5620-5624.
- Cavalli-Sforza (L.L.), Piazza (A.), Menozzi (P.), Mountain (J.) 1988, Reconstruction of human evolution: bringing together genetic, archaeological, and linguistic data, *Proceedings of the National Academy of Sciences, USA*, 85: 6002-6006.
- Chakraborty (R.) 1976, Cultural, language and geographical correlates of genetic variability in Andean highland Indians, *Nature* 264: 350-352.
- Chen (J.), Sokal (R.R.), Ruhlen (M.) 1995, Worldwide analysis of genetic and linguistic relationships of human populations, *Human Biology* 67: 595-612.
- Clouse (D.A.) 1996, Towards a reconstruction and reclassification of the Lakes Plains languages of Irian Jaya, in K.J. Franklin (ed.), *Papers in Papuan Linguistics 2; Pacific Linguistics A*, 85, Research school of Pacific and Asian studies, Australian National University, Canberra, p. 133-236.
- Darwin (C.) 1859a, *On the Origin of Species "L'origine des Espèces"*, Traduction de l'édition anglaise définitive par Barbier, 1882, Reinwald Ed.
- Darwin (C.) 1859b, *On the Origin of Species*, Bantam Books (1999), New York.
- Darwin (C.) 1871, *The descent of Man, "La descendance de l'Homme"*, Traduction de la deuxième édition anglaise par Barbier, 1891, Reinwald Ed.
- Dobzhansky (T.) 1966, *L'homme en évolution*, Flammarion, Paris.
- Dupanloup de Ceunick (I.), Schneider (S.), Langaney (A.), Excoffier (E.) 2000, Inferring the impact of linguistic boundaries on population differentiation: application to the Afro-Asiatic-Indo-European case, *European Journal of Human Genetics* 8: 750-756.
- Dyen (I.) 1962, The lexicostatistical classification of the Malayopolynesian languages, *Language* 38: 38-46.
- Dyen (I.), Kruskal (J.B.), Black (P.) 1992, *An Indoeuropean Classification: A lexicostatistical experiment*, The American Philosophical Society, Philadelphia.
- Ellegard (A.) 1959, Statistical measurement of linguistic relationship, *Language* 35: 131-156.
- Evrard (E.) 1966, in *Statistique et Analyse linguistique, Colloque de Strasbourg 20-22 avril 1964*, Presses Universitaires de France, Paris, p. 85-94.
- Farris (J.S.) 1972, Estimating phylogenetic trees from distance matrices, *American Naturalist* 106: 645-668.
- Farris (J.S.) 1981, Distance data in phylogenetic analysis, in V.A. Funk, D.R. Brooks (eds), *Advances in Cladistics*, New York Botanical Garden, New York, p. 3-23.
- Farris (J.S.) 1985, Distance data revisited, *Cladistics* 1: 67-85.

- Farris (J.S.) 1986, On the boundaries of phylogenetic systematics, *Cladistics* 2: 14-27.
- Felsenstein (J.) 1984, Distance methods for inferring phylogenies: a justification, *Evolution* 38: 16-24.
- Felsenstein (J.) 1986, Distance methods: a reply to Farris, *Cladistics* 2: 130-143.
- Fitch (W.M.) 1984, Cladistics and other methods: problems, pitfalls and potentials, in T. Duncan, T.F. Stuessy (eds), *Cladistics: perspectives on the reconstruction of evolutionary history*, Columbia University Press, New York, p. 221-252.
- Forster (P.), Renfrew (C.) 2006, *Phylogenetic Methods and the Prehistory of Languages*, McDonald Institute for Archaeological Research, University of Cambridge, Cambridge.
- Gaillard-Corvaglia (A.), Léonard (J.L.), Darlu (P.) 2007, Testing cladistics on dialect networks and phyla (gallo-romance and southern italo-romance), in Linguistics, T.a.f.C. (ed.), Ninth meeting of the ACL Special Interest Group in Computational Morphology and Phonology, Prague, p. 23-30.
- Goebel (H.) 1981, Éléments d'analyse dialectométrique (avec application à l'AIS), *Revue de Linguistique Romane* 45 : 349-420.
- Goebel (H.) 1983, Parquet polygonal et treillis triangulaire, Les deux versants de la dialectométrie interponctuelle, *Revue de Linguistique Romane* 47 : 353-412.
- Goebel (H.) 2004, Sprache, Sprecher und Raum: Eine kurze Darstellung der Dialektometrie, *Mitteilungen der Österreichischen Geographischen Gesellschaft* 146: 247-286.
- Gordon (R.J.) 2005, *Ethnologue: Languages of the world*, Dallas, Texas.
- Gray (R.D.), Atkinson (Q.D.) 2003, Language-tree divergence times support the Anatolian theory of Indo-European origin, *Nature* 426: 435-439.
- Gray (R.D.), Jordan (F.M.) 2000, Language trees support the express-train sequence of Austronesian expansion, *Nature* 405: 1052-1055.
- Greenberg (J.H.) 1960, The general classification of Central and South American languages, in A. Wallace (ed.), *Selected papers of the Fifth International Congress of Anthropological and Ethnological Sciences*, University of Pennsylvania Press, Philadelphia, p. 791-794.
- Greenberg (J.H.) 1987, *Language in the Americas*, Stanford University Press, Stanford.
- Greenhill (S.J.) 2002, *Testing the Express Train: Models of language evolution and human colonisation of the Pacific*, Unpublished Honours Dissertation, Department of Psychology, University of Auckland, Auckland, New-Zealand.
- Greenhill (S.J.), Gray (R.D.) 2005, Testing population dispersal hypotheses: Pacific settlement, phylogenetic trees and austronesian languages, in R. Mace, C.J. Holden, S. Shennan (eds), *The Evolution of Cultural Diversity: Phylogenetic Approaches*, Routledge Cavendish Publishers, p. 31-52.
- Haeckel (E.) 1874, *Heinr Anthropogenie oder Entwicklungsgeschichte*, "Anthropogénie ou Histoire de l'évolution humaine", Traduction de la deuxième édition allemande par Letourneau, 1877, Reinwald Ed.
- Hennig (W.) 1950, *Grundzüge einer Theorie der phylogenetischen Systematik*, Deutscher Zentralverlag, Berlin.
- Hoenigswald (H.M.), Wiener (L.F.) 1987, *Biological Metaphor and Cladistic Classification: an Interdisciplinary Perspective*, University of Philadelphia Press, Philadelphia.

- Holden (C.J.) 2002, Bantu language trees reflect the spread of farming across sub-Saharan Africa: a maximum-parsimony analysis, *Proceedings of the Royal Society B London*, 269: 793-799.
- Hunley (K.), Long (J.C.) 2005, Gene flow across linguistic boundaries in Native North American populations, *Proceedings of the National Academy of Sciences, USA*, 102: 1312-1317.
- Kessler (R.) 1995, Computational dialectology in Irish Gaelic, in *Seventh Conference of the European Chapter of the Association for Computational Linguistics*, ACL Publisher, Dublin, p. 60-66.
- Kroeber (A.L.) 1960, Statistics, Indo-European and Taxonomy, *Language* 36: 1-21.
- Laks (B.) 2002, Le comparatisme : de la généalogie à la génétique, *Languages* 146 : 19-46.
- Laks (B.) 2006, *Origin and Evolution of Languages: Approaches, Models, Paradigms*, Equinox Publishing, Londres.
- Lass (R.) 1997, *Historical Linguistics and Language Change*, Cambridge University Press, Cambridge.
- Lum (J.K.), Cann (R.L.), Martinson (J.J.), Jorde (L.B.) 1998, Mitochondrial and nuclear genetic relationships among Pacific Island and Asian populations, *American Journal of Human Genetics* 63: 613-624.
- Manni (F.), Guerard (E.), Heyer (E.) 2004, Geographic patterns of (genetic, morphologic, linguistic) variation: how barriers can be detected by using Monmonier's algorithm, *Human Biology* 76: 173-190.
- Michener (C.D.), Sokal (R.R.) 1957, A quantitative approach to a problem in classification, *Evolution* 11: 130-162.
- Nakhleh (L.), Ringe (D.), Warnow (T.) 2005, Perfect phylogenetic networks: a new methodology for reconstructing the evolutionary history of natural languages, *Journal of the Linguistic Society of America* 81: 382-420.
- Nerbonne (J.), Heeringa (W.), Kleiweg (P.) 1999, Edit distance and dialect proximity, in S.D.a.K. Joseph (ed.), *Time Warps, String Edits and Macromolecules: the Theory and Practice of Sequence Comparison*, CSLI Press, Stanford, p. v-xv.
- O'Grady (R.T.), Goddard (I.), Bateman (R.M.), DiMichele (W.A.), Funk (V.A.), Kress (W.J.), Mooi (R.), Cannell (P.F.) 1989, Genes and tongues, *Science* 243: 1651.
- Osthoff (H.), Brugmann (K.) 1878, *Morphologische Untersuchungen auf dem Gebiete der indogermanischen Sprachen*, Hirzel, Leipzig.
- Poloni (E.S.), Semino (O.), Passarino (G.), Santachiara-Benerecetti (A.S.), Dupanloup (I.), Langaney (A.), Excoffier (L.) 1997, Human genetic affinities for Y-chromosome P49a,f/TaqI haplotypes show strong correspondence with linguistics, *American Journal of Human Genetics* 61: 1015-1035.
- Quatrefages A. (de) 1886, « Introduction à l'étude des Races Humaines », Reinwald Ed.
- Ringe (D.) 1999, Language classification: scientific and unscientific methods, in B. Sykes (ed.), *The Human Inheritance: Genes, Language and Evolution*, Oxford University Press, Oxford, p. 45-74.
- Ringe (D.), Taylor (A.), Warnow (T.) 2002, Indo-European and computational cladistics, *Transactions of the Philological Society* 100: 59-129.

Rosser (Z.H.), Zerjal (T.), Hurles (M.E.), Adojaan (M.), Alavantic (D.), Amorim (A.), Amos (W.), Armenteros (M.), Arroyo (E.), Barbujani (G.), Beckman (G.), Beckman (L.), Bertranpetit (J.), Bosch (E.), Bradley (D.G.), Brede (G.), Cooper (G.), Corte-Real (H.B.), Knijff (de) (P.), Decorte (R.), Dubrova (Y.E.), Evgrafov (O.), Gilissen (A.), Glisic (S.), Golge (M.), Hill (E.W.), Jeziorowska (A.), Kalaydjieva (L.), Kayser (M.), Kivisild (T.), Kravchenko (S.A.), Krumina (A.), Kucinskas (V.), Lavinha (J.), Livshits (L.A.), Malaspina (P.), Maria (S.), McElreavey (K.), Meitinger (T.A.), Mikelsaar (A.V.), Mitchell (R.J.), Nafa (K.), Nicholson (J.), Norby (S.), Pandya (A.), Parik (J.), Patsalis (P.C.), Pereira (L.), Peterlin (B.), Pielberg (G.), Prata (M.J.), Previdere (C.), Roewer (L.), Rootsi (S.), Rubinsztein (D.C.), Saillard (J.), Santos (F.R.), Stefanescu (G.), Sykes (B.C.), Tolun (A.), Villems (R.), Tyler-Smith (C.), Jobling (M.A.) 2000, Y-chromosomal diversity in Europe is clinal and influenced primarily by geography, rather than by language, *American Journal of Human Genetics* 67: 1526-1543.

Ruhlen (M.) 1976, *A Guide to the Languages of the World*, M. Ruhlen Editor, Stanford.

Ruhlen (M.) 1987, *A Guide to the World's Languages*, Stanford University Press, Stanford.

Ruhlen (M.) 1994, *The Origin of Language, Tracing the Evolution of the Mother Tongue*, John Wiley and Sons, New York.

Saitou (N.), Nei (M.) 1987, The neighbor-joining methods: a new method for reconstructing trees, *Molecular Biology and Evolution* 4: 406-425.

Schleicher (A.) 1853, Die ersten Spaltungen des indogermanischen Urvolkes, *Allgemeine Zeitung fuer Wissenschaft und Literatur*: 786-787.

Schleicher (A.) 1863, Die Darwinsche Theorie und die Sprachwissenschaft -offenes Sendschreiben an Herrn Dr. Ernst Haeckel. Weimar, H. Boehlau.

Sneath (P.H.A.), Sokal (R.R.) 1973, *Numerical Taxonomy*, W.H. Freeman, San Francisco.

Sokal (R.R.) 1988, Genetic, geographic, and linguistic distances in Europe, *Proceedings of the National Academy of Sciences, USA*, 85: 1722-1726.

Sokal (R.R.), Sneath (P.H.A.) 1963, *Principles of Numerical Taxonomy*, W.H. Freeman, San Francisco.

Swadesh (M.) 1952, Lexico-statistic dating of prehistoric ethnic contacts, *Proceedings of the American Philosophical Society* 96: 453-463.

Szulmajster-Celnikier (A.) 2006, La question de l'origine des langues : vaine quête du Graal ? *Marges linguistiques* 11: 327-342.

Tort (P.) 1980, *Evolutionisme et linguistique*, Vrin, Paris.

Trask (R.L.) 1996, *Historical Linguistics*, Oxford University Press, Oxford.

Wood (E.T.), Stover (D.A.), Ehret (C.), Destro-Bisol (G.), Spedini (G.), McLeod (H.), Louie (L.), Bamshad (M.), Strassmann (B.), Soodyall (H.), Hammer (M.F.) 2005, Contrasting patterns of Y chromosome and mtDNA variation in Africa: evidence for sex-biased demographic processes, *European Journal of Human Genetics* 13: 867-876.

## NOTES

1. Ce texte a été présenté lors d'un séminaire de l'EHESS sur « Comportement, représentation, culture », Collège de France, le 4 avril 2002. Depuis cette date, les

organisateur du séminaire n'ayant pu assurer leur projet de publication, le présent texte a été réactualisé.

---

## RÉSUMÉS

Ce texte propose un point de vue critique sur la façon dont l'histoire des langues et l'histoire des populations humaines se sont construites en parallèle et en correspondance, depuis le début du XIXe s. jusqu'à nos jours. D'un point de vue méthodologique, les convergences sont nombreuses, les linguistes, aussi bien que les généticiens des populations utilisant les mêmes méthodes de distances, ou les mêmes approches cladistiques ou probabilistes de reconstruction phylogénétique.

La validité et les limites des analogies entre évolution linguistique et génétique sont discutées à partir d'exemples tirés de la littérature récente. L'analyse critique se place selon un double point de vue : d'abord en examinant comment ces diverses méthodes sont appliquées par les phylogénéticiens aux données linguistiques, sous le regard suspicieux de certains linguistes et, ensuite, en considérant la façon dont les généticiens des populations intègrent certaines classifications linguistiques, souvent contestées par les linguistes, à leurs propres problématiques génétiques.

This text is intended to review how the history of languages and the history of human populations were drawn in parallel and in correspondence, since the 19th century. From the methodological point of view, convergences are numerous, both linguists and population geneticists making use of the same distance methods and the same cladistic or probabilistic approaches.

The validity and limits of this analogy between linguistic and genetic evolution of mankind are discussed from recent published examples. Critics are based alternately on how these methods are applied to the linguistic data by phylogeneticists, under the suspicious and critical look of some linguists, and on how the human population geneticists include linguistic classification, often much debated by linguists, within their own genetic concerns.

## INDEX

**Keywords** : evolution, genetics, linguistics, methodology

**Mots-clés** : évolution, génétique, linguistique, méthodologie

## AUTEURS

**MAHÉ BEN HAMED**

UMR 5596, Laboratoire de Dynamique du langage, Institut des Sciences de l'Homme, 14 avenue Berthelot, 69363 Lyon CEDEX 07, France, e-mail : Mahe.Ben-Hamed@ish-lyon.cnrs.fr.

**PIERRE DARLU**

INSERM U535, CNRS, Génétique épidémiologique et structure des populations humaines, Hôpital Paul Brousse, BP 1000, 94817 Villejuif CEDEX, France, e-mail : [pierre.darlu@inserm.fr](mailto:pierre.darlu@inserm.fr) et Université Paris-Sud, UMR-S535, 94817 Villejuif CEDEX, France.