Evaluation de la congruence entre évolutions génétique, morphologique et linguistique : modèles et méthodes

DARLU Pierre

Action « Origine de l'Homme, du Langage et des Langues »

A. FICHE ADMINISTRATIVE

Titre du projet :

Evaluation de la congruence entre évolutions génétique, morphologique et linguistique : modèles et méthodes

Mots-clés :
Evolution, langues, génétique, congruence

Résumé du projet (10 lignes maximum) :
Ce projet se propose d'analyser, comparer et évaluer les différents *modèles* utilisés pour la reconstruction de l'histoire des langues et des populations. L'accent sera mis sur la notion de *caractère*, génétique et linguistique, abordée dans une perspective multivariée, sur les hypothèses concernant leur mode d'évolution, sur l'importance accordée à chacun d'eux, sur la définition de *critères* qui permettent des regroupements en classe de ressemblance (des populations et des langues), sur les *méthodes* qui conduisent à dresser des filiations historiques entre ces classes. Différentes méthodes permettant de tester la congruence entre histoire des gènes et histoire des populations seront testées et appliquées sur plusieurs exemples tirées d'expérience de terrain à différentes échelles (Nouvelle Guinée, Tibéto-birman, langues romanes, maya, etc.). Dans une étape ultérieure, l'étude de congruence sera étendue aux descriptions morphologiques des populations actuelles et passées.

1. Responsable scientifique du projet

Nom ...DARLU..... Prénom.....Pierre.....……
Grade.....DR2 (CNRS).....……...
Discipline du responsable scientifique: ...Génétique, Phylogénie .....……..
Établissement de rattachement .INSERM U535, affiliée CNRS.....……
.....……
Adresse professionnelle : N°, rue , BP ...80, Rue du Général Leclerc.....……
.....……
Code postal 94276 Commune .....Le Kremlin Bicêtre...…….

Tél 01 49 59 53 40 Fax 01 49 59 53 31
E-Mail : …darlu@kb.inserm.fr

2. Laboratoire ou organisme de rattachement de l'équipe de recherche

Intitulé ...INSERM U535 Génétique épidémiologique et Structure des populations humaines
Type de formation *(cocher la case utile)*
- *Unités CNRS :* ☐ unité propre du CNRS ☐ unité associée ou mixte du CNRS
  Préciser le code unité .....
  Préciser la délégation régionale : .....
- *Unités hors CNRS :* ☐ unité universitaire *(Préciser l'université)*
  .....…
  ☒ Autre *(Unité INSERM affiliée CNRS)*

Nom du directeur de l'organisme : .....Mr Griscelli.....
Adresse : N° , rue , BP.....101, rue de Tolbiac.....…

.....
Code postal 75013 Commune .....Paris.....
Tél I_ı_ı_ı_ı_ı_ı_ı_ı_ı_I Fax I_ı_ı_ı_ı_ı_ı_ı_ı_ı_I

3. Autre(s) laboratoires ou organisme(s) partenaires

Intitulé: CNRS / LACITO

Discipline(s) couverte(s) par l'équipe : Linguistique, ethnologie, musicologie

Nom du directeur de l'équipe : Mme Zlatka GUENTCHEVA

Membre(s) partenaire(s) :

Nom : Mazaudon Prénom : Martine
Grade : DR2 E-Mail : mazaudon@vjf.cnrs.fr
Nom : Jacobson Prénom : Michel
Grade : IE2 E-Mail : jacobson@idf.ext.jussieu.fr

Adresse : CNRS/LACITO Centre André-Georges Haudricourt, 7 rue Guy Môquet, Bât. 23

Code postal 94801 Commune Villejuif Cedex

Tél 01 49 58 37 52/56 Fax 01 49 58 37 79

3. Autre(s) laboratoires ou organisme(s) partenaires

Cette section est à reproduire autant de fois que nécessaire.
Intitulé ..Laboratoire de Dynamique du Language, CNRS
Discipline(s) couverte(s) par l'équipe : Linguistique historique

Nom du directeur de l'équipe :.J.M. Hombert

Membre(s) partenaire(s) :

Nom : …MARSICO……..Prénom : …Egidio……
Grade : Doctorant……
E-Mail : Egidio.Marsico@ish lyon.cnrs.fr

Adresse : N° 14, avenue Berthelot 69363 Lyon cedex 07
Code postal 69363 Commune Lyon Cedex 07
Tél 04 72 72 64 62 Fax 04 72 72 65 90

3. Autre(s) laboratoires ou organisme(s) partenaires

Intitulé CNRS-IMAG Laboratoire Leibniz Equipe Laplace
Discipline(s) couverte(s) par l'équipe : Modèles aléatoires en robotique et intelligence artificielle

Nom du directeur de l'équipe : Pierre Bessière

Membre(s) partenaire(s) :

Nom : Bessière Prénom : Pierre
Grade : CR1 E-Mail : pierre.bessiere@imag.fr

Adresse : CNRS-IMAG/Leibniz , 40 avenue Félix Viallet
Code postal 38031 Commune Grenoble

Tél 04 76 57 46 73 Fax 04 76 57 46 023. Autre(s) laboratoires ou organisme(s) partenaires

Nom ...DEGIOANNI..... Prénom.....Anna.....……

Grade...Postdoc.....……..

Discipline du responsable scientifique: ...Génétique des populations.....……..

Établissement de rattachement .INSERM U535, affiliée CNRS.....……

Adresse professionnelle : N°, rue , BP ...80, Rue du Général Leclerc.....…….

Code postal _94 276 Commune .....Le Kremlin Bicêtre...…….

Tél 01 49 59 53 39 Fax 01 49 59 53 31
E-Mail : …degio@kb.inserm.fr

. Autre(s) laboratoires ou organisme(s) partenaires

Intitulé: Institut de la Communication Parlée / Structure du code
Discipline(s) couverte(s) par l'équipe : Universaux phonétiques, études sur le lexique, acquisition production de la parole…

Nom du directeur de l'équipe : Pierre Escudier

Membre(s) partenaire(s) :

Nom : Boë Prénom : Louis-Jean
Grade : IR1 E-Mail : boe@icp.inpg.fr
Nom : Vallée Prénom : Nathalie
Grade : CR2 E-Mail : vallee@icp.inpg.fr

Adresse : ICP Université Stendhal BP 25
Code postal 38031 Commune Grenoble Cedex 9

Tél 04 76 82 43 38 Fax 04 76 82 43 35

3. Autre(s) laboratoires ou organisme(s) partenaires

Intitulé: Université de Sydney / Département de linguistique

Discipline(s) couverte(s) par l'équipe : Sciences du langage

Nom du directeur de l'équipe :

Membre(s) partenaire(s) :
Nom : Foley Prénom : William
Grade : professeur E-Mail : william.foley@linguistics.usyd.edu.au

Adresse : University of Sydney, New South Wales 2006 AUSTRALIE
Code postal Commune

Tél Fax

3. Autre(s) laboratoires ou organisme(s) partenaires

Intitulé Institut de Mathématiques de Luminy (IML)
Discipline(s) couverte(s) par l'équipe : Mathématiques, Informatique
Nom du directeur de l'équipe : Brigitte Mosse
Membre(s) partenaire(s) :

Nom : Guénoche Prénom : Alain
Grade : CR1 E-Mail : Guenoche@iml.univ-mrs.fr

Adresse : IML 163, Avenue de Luminy,
Code postal 13288 Commune Marseille Cedex 9

Tél 04 91 26 85 63 Fax 04 91 26 96 55

3. Autre(s) laboratoires ou organisme(s) partenaires

Intitulé: Université de Californie à Berkeley / Département de linguistique
Discipline(s) couverte(s) par l'équipe : Sciences du Langage
Nom du directeur de l'équipe : président du département : Larry M. Hyman
Membre(s) partenaire(s) :
Nom : Matisoff Prénom : James A.
Grade : Professeur E-Mail : matisoff@socrates.berkeley.edu
Nom : Lowe Prénom : John Brandon
Grade : Assistant researcher E-Mail : jblowe@askjeeves.com
Adresse : Department of Linguistics, University of California, 1203 Dwinelle Hall, Berkeley, CA 94720-2650
Code postal 94720-2650 Commune Berkeley CA
Tél Fax

3. Autre(s) laboratoires ou organisme(s) partenaires

Intitulé: Université de Sydney/ Département d'informatique
Discipline(s) couverte(s) par l'équipe :
Nom du directeur de l'équipe :
Membre(s) partenaire(s) :
Nom : Patrick Prénom : Jon
Grade : professeur E-Mail : jonpat@cs.usyd.edu.au

Adresse : Basser Department of Computer Science, Madsen Building F09, University of Sydney, New South Wales 2006 AUSTRALIE

3. Autre(s) laboratoires ou organisme(s) partenaires

Intitulé Université de Paris III
Discipline(s) couverte(s) par l'équipe : Linguistique et phonétique générales et appliquées
Nom du directeur de l'équipe : administrateur provisoire: Etienne Pietri
Membre(s) partenaire(s) :
Nom : Rebuschi Prénom : Georges
Grade : professeur E-Mail : rebuschi@idf.ext.jussieu.fr

3. Autre(s) laboratoires ou organisme(s) partenaires

Futur équipe : LACITO en 2001
Nom : Léonard Prénom : Jean-Léo
Grade : maître de conférence E-Mail : leonard@idf.ext.jussieu.fr

3. Autre(s) laboratoires ou organisme(s) partenaires

Futur équipe : Osterlits CNRS
Adresse : ILPGA Université de Paris III, 19 rue des Bernardins
Code postal 75005 Commune Paris

B. PROJET SCIENTIFIQUE

Evaluation de la congruence entre évolutions génétique, morphologique et linguistique : modèles et méthodes

La grande diversité de l'espèce humaine peut s'appréhender à différents niveaux : *morphologique* d'abord, à la façon des anthropologues physiques du XIXème et du début du XXème siècle et comme le font de nos jours les paléo-anthropologues sur les restes osseux, *génétique* ensuite, dans la tradition de la génétique des populations en se fondant sur l'étude des polymorphismes classiques (ABO, Rh, Gm, HLA…) ou des polymorphismes de l'ADN, *culturel et linguistique* enfin comme de nombreux auteurs l'ont fait, depuis Schlegel au début du XIXème siècle jusqu'à Greenberg, Ruhlen et leurs contradicteurs, de nos jours.

Ces trois directions de recherche ont en commun la préoccupation de définir des *critères* qui permettent des regroupements en classe de ressemblance d'une part et de mettre en place des *méthodes* qui conduisent à dresser des filiations historiques entre ces classes d'autre part. En revanche elles diffèrent à la fois par la nature des critères retenus, par les hypothèses posées sur leur mode d'évolution et par l'importance accordée à chacun d'eux.

Ces recherches conduisent naturellement à établir des relations entre les apparentements linguistiques et les apparentements inférés par la génétique des populations. Cette problématique a été mise à l'honneur à la suite des travaux récents de Greenberg (1987) et Ruhlen (1975,1994, 1997), pour la partie linguistique, et par les équipes de L.L. Cavalli-Sforza (1988) et de Sokal (1988) et Chen et al. (1995), pour la partie génétique.

Cependant, malgré les évidents succès de ces travaux comparatifs sur l'histoire des langues et l'histoire des gènes, on ne peut pas considérer que les problèmes de méthodes soient totalement résolus. *Les modèles ou métaphores* qui expriment les conceptions des apparentements dans les champs de la linguistique et de la génétique ne sont que des *approximations qui doivent être évaluées et font appel à des méthodes de reconstruction qui méritent d'être adaptées.*

Par ailleurs, il est primordial de distinguer entre l'éventuelle analogie des méthodes employées pour reconstruire l'histoire des populations et celle des langues d'une part et les " coïncidences " que l'on peut remarquer dans les résultats de ces reconstructions d'autre part. Bien des travaux confortent les regroupements de populations basés sur des critères génétiques par leur coïncidence avec des classifications en familles ou sous-familles de langues telles qu'elles sont préconisées par certains linguistes. La tentation est alors grande de vouloir valider l'histoire des langues par l'histoire génétique des populations qui parlent ces langues et réciproquement.

Dans ce domaine, l'application de méthodes rigoureuses d'évaluation de la congruence entre histoires évolutives différentes s'impose donc.

Le présent projet se propose donc de répondre à plusieurs de ces interrogations pour lesquelles réflexions sur les modèles et recherches sur les méthodes se juxtaposent, et de l'illustrer par quelques exemples :

1 - Modèles d'évolution et méthodes de reconstruction phylogénétique. Les modèles d'évolution génétique sont-ils applicables à l'évolution des langues ou comment les adapter? Cette partie comprend plusieurs axes :
1.1.- Le modèle de l'arbre
1.2 - Les caractères et la méthode cladistique
1.3 - Les approches statistiques et probabilistes
1.4 - Les vitesses d'évolution et le problème des datations

2 - La congruence : Les conclusions obtenues séparément sur l'histoire évolutive des langues et celle des populations, définies morphologiquement ou génétiquement, sont-elles congruentes ?

3 - Exemples, pris à différentes échelles géographiques et différentes profondeurs historiques.

1 - Modèles d'évolution et Méthodes de reconstruction phylogénétique :

1.1 - Le modèle de l'arbre

Sur le plan des modèles de diversification, le modèle standard de l'évolution biologique est l'arbre. Cependant, ce modèle de l'arbre reste problématique quand il est appliqué à la diversification entre populations d'une même espèce. En effet, comme les populations ne sont pas des unités reproductives isolées, et peuvent donc échanger des conjoints, le modèle dichotomique de l'arbre n'est plus satisfait. Les autres modèles alternatifs reposent sur la notion de différentiation génétique avec la distance (voir les travaux de Malécot, Kimura, Morton). Cependant, plusieurs méthodes permettent de traiter simultanément la différentiation temporelle sous forme d'arbre tout en intégrant la possibilité de mélanges à différent niveau de l'arbre (Cavalli-Sforza et Piazza, 1975, Lathrop, 1982, Bertorelle et al..).

De même, en linguistique au moins deux modèles sont en compétition, l'arbre et les vagues, sans compter les variations qui combinent ces images (comme l'arbre à greffons!).

Une étude des métaphores utiles et moins utiles dans les deux domaines est à poursuivre sur le plan théorique et mathématique (collaboration avec Alain Guénoche) et sur le plan épistémologique et philosophique.

Les reconstructions de l'histoire évolutive des gènes et des populations qui les portent sont fondées sur différentes classes de méthodes. L'une d'elles la méthode " phénétique ", analyse les ressemblances globales entre populations. Ces ressemblances sont quantifiées au travers de matrices de distances. La recherche des relations phylogénétiques entre les populations se base sur cette matrice et utilise des algorithmes d'agglomération (UPGMA, NJ) ou de recherche de vraisemblance maximale quand un modèle probabiliste d'évolution est possible. Dans ses comparaisons avec les données linguistiques, Cavalli-Sforza semble privilégier ce type de méthodes. En linguistique, quelques précurseurs ont proposé la mise en place de telles méthodes quantitatives (Evrard, 1966), non sans prudence quant aux conclusions sur la validité des parentés ainsi suggérées. Ces méthodes, très courantes, bien que souvent contestées, en génétique évolutive, sont probablement peu adaptées en linguistique. Cependant, des distances phonologiques, grammaticales ou sémantiques peuvent être éventuellement estimées et donc être traitées par ces différentes méthodes phénétiques, tout en s'affranchissant d'hypothèses lourdes comme celle d'un taux d'évolution constant (voir 1.4). Nous nous proposons de discuter et d'évaluer la pertinence d'une telle approche dans un contexte linguistique.

1.2 - Les caractères : définitions et méthodes

Parmi toutes les méthodes de reconstruction, la méthode cladistique est certainement la plus appropriée pour résoudre les problèmes de parenté, qu'ils soient biologiques ou linguistiques, dans la mesure où elle est fondée sur l'observation des caractères. Elle constitue d'ailleurs, sous le nom de principe des innovations partagées la pratique standard en linguistique historique (voir également Hoenigswald et Wiener, 1987). Elle nécessite, en préalable, une présentation explicite des hypothèses ou des modèles sous-jacents à l'évolution des caractères biologiques ou linguistiques. En particulier elle impose de proposer des définitions précises sur :
i) les caractères eux-mêmes,

ii) les différentes formes qu'ils peuvent prendre,
iii) les transformations ou transitions possibles entre ces formes,
iv) les relations entre caractères
v) le poids respectif à donner à chacun d'eux (par exemple, dans les langues du Nord de la Nouvelle-Guinée, la correspondance d'une morphologie supplétive est beaucoup plus informative que d'autres caractères).

Nous nous proposons d'élaborer une réflexion commune sur tous ces aspects entourant la notion de caractère en linguistique et ses analogies en biologie et génétique.

La méthode cladistique consiste à prendre en compte l'ensemble des hypothèses formulées sur les caractères, puis à construire la relation arborescente la plus parcimonieuse liant les gènes entre eux, ou les langues entre elles. Dans un modèle cladistique classique, le critère de parcimonie revient à minimiser le nombre d'hypothèses ad hoc, c'est-à-dire à minimiser les changements de type réversion ou convergence, sans inférence sur leurs causes (« aléatoire » « adaptative, emprunt… »).

Dans le contexte biologique portant sur le polymorphisme de l'ADN, une telle démarche qui aboutit à une représentation sous forme d'arbre trouve tout son sens, dans la mesure où les différentes formes génétiques peuvent être reliées entre elles par des mutations identifiables et dont la probabilité est généralement faible.

Dans le domaine linguistique par contre, étant donné les restrictions existant sur les types de mutation possibles, il n'est pas rare de rencontrer les mêmes traits à l'issue d'un cycle évolutif, en phonologie comme en syntaxe (voir le passage ergatif - absolutif => nominatif - accusatif => ergatif - absolutif attesté par l'histoire du persan; ou l'évolution e>ei>oj>we>e>ej rencontrée dans les parlers d'Oïl, avec le retour du phonème /e/ et la répétition de sa diphtongaison en /ej/). Nous poserons la question de formaliser dans une approche cladistique la notion de cycle évolutif.

Pour valider la méthode dans le domaine linguistique, il faut discuter la fréquence relative de divers événements, comme le changement, l'emprunt etc. et développer des approches spécifiques qui permettent de s'affranchir de la représentation simpliste en arbre, à l'image de ce qui se pratique dans les analyses de réseaux ou dans le domaine de la génétique (Barthélémy et Guénoche, 1991 ; Bandelt, 1995, Pritchard et al., 2000). Pour cela, nous nous appuierons également sur les travaux de Jon Patrick, associé au projet, qui portent sur les mesures de la parcimonie relative de reconstructions concurrentes.

1.3 - Des approches statistiques et/ou probabiliste

La méthode comparative classique et la classification cladistique qui en découle impliquent un progrès lent des connaissances. En effet le comparatiste professionnel tient toujours pour preuve d'une relation postulée l'histoire détaillée des étapes et événements qui relient la forme ancestrale reconstruite à ses descendants, suivant en cela Calvert Watkins (1990). Pour établir à tout le moins des hypothèses sur des classifications plus ambitieuses, divers auteurs ont proposé de relaxer les contraintes de la méthode comparative. En 1973 Benedict parlait de "téléo-reconstruction" en tibéto-birman, Greenberg (1987) embrasse des espaces encore plus vastes. Ruhlen (1994, 1997) renonce à la notion même de correspondance régulière pour ne plus rechercher que des traces, visibles, d'après lui, directement par des ressemblances de formes et de sens entre des éléments lexicaux dans différentes langues

A des profondeurs temporelles qui ne permettent plus, ou pas encore, l'enchaînement pas à pas des dérivations, en l'absence de l'histoire détaillée réclamée par Watkins, ou avant d'avoir pu l'établir, quelle foi peut-on apporter à des ressemblances apparentes de forme et de sens entre deux langues ? ou même à des correspondances ou corrélations occasionnelles?

Dans ce contexte, une évaluation statistique de la qualité des observables devient nécessaire, les travaux de Ringe ayant bien montré que la probabilité de ressemblances ou corrélations dues au hasard est loin d'être négligeable.

1.3.1 - Evaluation de la distance entre les traits phonologiques

Ringe propose une méthode mathématique pour tester, sur deux langues hypothétiquement apparentées, si le nombre de correspondances observées est statistiquement significatif (supérieur à ce qu'on attendrait du fait du seul hasard.). Le point de départ de la comparaison est la liste de 100 mots de Swadesh dans deux langues. Ringe sélectionne une variable particulière (par exemple la consonne initiale), construit une distribution théorique de cette variable en fonction de ses

fréquences d'apparition dans chacune des deux listes, et calcule la probabilité d'obtenir de manière aléatoire (loi binomiale) la distribution (le taux de correspondance) observée. Enfin il retient comme significatives les correspondances observées pour lesquelles cette probabilité est inférieure à 1%.

Ce travail pionnier a fait l'objet de critiques qui pourraient conduire à remettre en cause ses conclusions. Baxter et Manaster-Ramer (1996) critiquent l'aspect mathématique du travail et déclarent la méthode inadéquate entre autre du fait qu'elle ne tient pas compte des phénomènes de transphonologisations et d'autres évolutions dépassant le cadre du segment. Ringe lui-même a envisagé plusieurs modifications possibles de sa méthode, et a jugé dans un premier temps qu'elle ne changeaient pas les résultats de manière significative.

Nous reprendrons l'examen des modifications possibles, en commençant par un essai de définition plus fine des éléments phonologiques à comparer (le phonème pourrait être remplacé par le trait, et le calcul devrait se faire sur un mot entier, permettant la prise en compte de "ressemblances" éventuellement déplacées sur un segment contigu). Nous chercherons à établir pour les calculs une mesure de proximité phonétique qui pourrait s'inspirer au départ de la géométrie des traits de Clements.

Nous testerons les résultats obtenus à la lumière des relations connues dans les familles de langues étudiées par les différents participants (océanien, tibéto-birman, maya, langues romanes, finno-ougriennes, papoues, voir partie 3)

1.3.2.- Évaluation de la part des contraintes biologiques

Depuis les années 70, les études typologiques s'attachent à expliquer la présence de grandes tendances dans les systèmes phonologiques des langues du monde à partir des contraintes de production et de perception de parole basées 1) sur nos capacités à contrôler la position de la langue et celle de la mâchoire, 2) sur les capacités d'intégration des systèmes auditifs et visuels. Les modèles de prédiction des structures sonores basés sur ces contraintes, confrontés aux données typologiques, montrent que le choix des langues est en partie conditionné par les capacités articulatoires et perceptives : les voyelles favorisées par les langues sont celles qui ont une bonne forme perceptive (voyelles focales, cf. Schwartz, Boë, Vallée, Abry, 1997a&b), les consonnes et les structures syllabiques les plus fréquentes sont aussi celles du babillage et des premiers mots (Vallée, Boë, Stefanuto, 1999). On ne peut donc ignorer, dans une évaluation des ressemblances linguistiques, la probabilité beaucoup plus forte de certaines formes liées aux propriétés substantielles des unités sonores. Les résultats de ces recherches en cours seront intégrés au modèle.

1.3.3 - Evaluation de la distance sémantique

Un deuxième axe de cette recherche devrait être l'estimation de la faisabilité d'une mesure de la proximité sémantique. En effet Ringe se donne pour règle l'identité complète de signification entre les cognats potentiels tandis que Ruhlen s'autorise des libertés extrêmes. Des travaux comme ceux de Buck ou de Matisoff sur les glissements sémantiques attestés pourraient permettre d'encadrer le degré de variation acceptable. Nous partirons ici du travail très détaillé effectué sur les parties du corps en tibéto-birman pour tenter une modélisation sur un domaine limité. Les différents types de changements sémantiques y sont représentés dans les figurations par des tracés d'épaisseur ou de nature différente. L'attribution de "points" suivant la facilité du passage d'un sens à un autre sera tentée pour un calcul de la distance sémantique parcourue entre cognats potentiels.

1.4 - Vitesse d'évolution et le problème des datations

La question des vitesses de changement des caractères génétiques et linguistiques fait aussi l'objet de débats. En génétique, le problème est d'évaluer la taille efficace des populations et/ou les taux de mutations du matériel génétique. La datation des évènements de différentiation dépend des valeurs attribuées à ces paramètres qui peuvent varier dans le temps, d'une population à l'autre et d'un gène à l'autre. De même, en linguistique, après les hypothèses simplificatrice d'un taux de renouvellement supposé constant en glottochronologie (Swadesh, 1950, 1952), le taux de renouvellement lexical a été considéré comme variable à la fois à l'intérieur du lexique mais également d'une langue à l'autre, limitant ainsi sérieusement la possibilité de dater les nœuds des arbres proposés, et donc, les proto-langues reconstruites. En ce qui concerne la comparaison multilatérale, la principale critique est que les pourcentages de ressemblances utilisés pour identifier les liens historiques d'apparentement sont souvent équivalents à ce que donnerait le hasard (voir notamment Ringe, 1992, 1995). La question revient finalement à savoir si une trop longue période de divergence ne finit pas par effacer toutes traces d'apparentement, conduisant à déterminer un

seuil de différenciation à partir duquel une langue donnée cesse d'être au profit d'une autre.

D'autres modèles abordent le phénomène d'évolution des langues dans sa totalité. Dixon (1997), par exemple, propose un modèle dit de "ponctuations et d'équilibres" adapté du modèle de S. Jay Gould en biologie, dans lequel il pointe le fait important que le processus de divergence des langues (ponctuations), résumé par des représentations arborescentes "génétiques", n'englobe pas la totalité du phénomène de mutation des langues, et que de longues phases d'équilibre amènent des convergences typologiques entre les langues par suite de diffusion de traits lors des contacts.

Un autre modèle général est proposé par Nettle (1999), qui, sans se préoccuper du comment de la divergence des langues (but qu'il confère à la linguistique historique) essaie de voir pourquoi elles se diversifient, en adaptant partiellement le "neutral-mutation model" de Kimura, (1983), développé en évolution biologique. Dans cette optique, Nettle propose la notion de pool linguistique humain dans lequel les éléments atomiques ne seraient pas les langues mêmes, mais ce qu'il appelle des "items linguistiques".

Enfin, les méthodes fondées sur l'analyse des distributions des différences entre langues prises deux à deux (pairwise differences), courante en génétique des populations humaines dans le cadre des théories de la coalescence, pourraient être explorées dans le domaine de la linguistique, selon une analogie à évaluer, en particulier dans la recherche de chronologie et de modalité d'expansion des langues.

2 - La congruence

Une fois les histoires évolutives des populations et des langues établies, indépendamment les unes des autres, se pose le problème de tester leur congruence, c'est-à-dire de déterminer si ces deux "histoires" sont convergentes, en totalité ou partiellement, et d'évaluer statistiquement la puissance des tests et les risques d'erreur. Pour cela, différentes méthodes seront développées, à partir de celles déjà existantes : méthodes de rééchantillonnages (test ILD de Farris et al. (1995), comparaison d'arbres (critères de Robinson et Foulds (1981), recherche d'évolutions corrélées (Page, 1990 ; Harvey et Pagel, 1991). Comme la représentation en arbre n'est souvent qu'une description de relations dans l'espace, des tests seront développés pour évaluer si la représentation arborescente est significativement meilleure qu'une représentation euclidienne qui ne ferait que traduire une simple ressemblance (entre langues ou gènes) directement dépendant de la distance géographique qui les séparent. Des degrés supérieurs de relation peuvent également être testés (Guénoche, 2000). Les applications se feront à partir des différents domaines linguistiques détaillés dans le chapitre suivant 3, et après constitution de bases de données génétiques (polymorphisme classique, ADN mitochondrial ou chromosome Y) portant sur les mêmes aires géographiques, linguistiques ou culturelles.

3 - Exemples :

Les participants au projet, spécialistes de différentes régions du monde, mettront en commun les matériaux récoltés sur le terrain, leur connaissance et leur réflexion sur les questions évoquées plus haut, afin de les confronter avec les modèles et les méthodes élaborées dans les autres champs disciplinaires (méthodes statistiques et probabilistes, génétique des populations, phylogénie…)

3.1. Un cas d'école: les langues de Nouvelle Guinée

Deux grandes familles de langues se partagent cette région, la famille austronésienne sur les côtes, et la famille «trans-Nouvelle-Guinée» dans les terres, chacune avec environ 250 langues. En dehors de ces deux ensembles, on dénombre environ 500 langues qui appartiennent à une trentaine de petites familles de faible extension géographique, ou qui sont des isolats (environ deux douzaines). Ces autres familles ne contiennent pas plus d'une vingtaine de langues chacune, et leurs locuteurs vont de 2000 à une cinquantaine de personnes.

Cette situation complexe dure certainement depuis des millénaires, avec, dans certains cas, un multilinguisme ancien et des influences structurelles mutuelles profondes entre les langues. L'application de la méthode comparative classique pour la détermination des relations génétiques entre les langues s'en trouve fortement complexifiée. Des emprunts de toute nature sont attestés, y compris dans le vocabulaire de base (parties du corps, noms de parenté, petits nombres, pronoms…) généralement réputé peu susceptible d'emprunt. Il est donc risqué de se fier ici à des cognats apparents dans le vocabulaire de base. L'emprunt de schémas structurels y est aussi manifeste..

Dans cette situation, les meilleurs marqueurs d'une parenté génétique demeurent sans doute les traces de morphologie irrégulière et supplétive, la complexité de ces systèmes rendant improbable une convergence due au hasard ou un emprunt. Le cas des langues de la famille Sepik-Ramu est particulièrement intéressant. On considère en effet qu'en dépit de l'absence quasi-totale de cognats lexicaux entre les deux sous-familles de la Basse Sepik et la Basse Ramu, on peut démontrer leur parenté sur la base d'un certain nombre de traits de morphologie. Ce type de rapports nous amènera à incorporer dans nos efforts de modélisation une pondération différentielle de la valeur des critères d'apparentement suivant la nature des « ressemblances » mises en évidence.

D'autre part si cette hypothèse d'un apparentement presque sans cognats est acceptée, elle pose, par extrapolation, la question théorique de l'origine des isolats : seul survivant d'une famille ou résultat d'une divergence extrême ?

Dans le cadre du présent projet, nous concentrerons notre attention sur les basses-terres du Nord de la Nouvelle-Guinée, du bassin de la Mamberano au bassin de la Sepik et de la Ramu. Cette région représente une zone 'résiduelle' dans les termes de J. Nichols, et elle contient à peu près la moitié de toutes les langues de la région, et presque tous les isolats postulés y sont concentrés.

Sur la base de matériaux de terrain nous établirons des hypothèses plus concrètes sur les relations entre les langues de cette région et y testerons la plausibilité et l'adéquation de différents modèles d'apparentements linguistiques. Nous y travaillerons en collaboration étroite avec des collègues généticiens spécialistes des populations du Pacifique, en vue de tester la congruence des schémas obtenus dans les deux domaines.

3.2 Autres domaines

Le domaine des langues romanes et celui des langues tamang du Népal, nous offrirons, à l'opposé, des exemples de différenciation à faible profondeur temporelle, tandis que les langues maya constituent un groupement au niveau d'une famille élargie. Les langues finno-ougriennes pour leur part se sont éparpillées au point de présenter une 'insularisation' des parlers. Nous utiliserons ces situations différentes pour affiner les notions de traces, d'innovations partagées, ou d'emprunt à différents niveaux de ce qui restera de l'arbre généalogique.

3.2.1 Langues romanes : non seulement cette famille est relativement compacte (hormis le daco-roman), mais elle fait partie des familles les mieux étudiées et les plus documentées au monde par les archives philologiques. Il existe en outre un nombre considérable d'atlas linguistiques, dont l'ALR (= Atlas Linguistique Roman, dirigé par Michel Contini, Grenoble), qui permet d'étudier l'aréologie et les mécanismes de diffusion des variables typologiques, les phénomènes de conservation et d'innovation structurale. Malgré cette situation privilégiée, la classification des langues et dialectes romans pose des problèmes très sérieux aux typologues et aux classificateurs (validité des critères : phonologiques, morphologiques, lexicaux, etc.). Nous utiliserons cette famille comme test pour la modélisation de la cyclicité des changements, et pour une appréciation de la "naturalité" des changements (en relation avec le point 1.3.2).

3.2.1. Langues tibéto-birmanes: la famille est riche de quelques 300 langues, qui ont fait l'objet d'efforts de classification systématique (avec développement de nomenclature spécifique pour les différents niveaux d'apparentements par R. Shafer dans les années '50), et en même temps de tentatives d'application de la "téléo-reconstruction". A un niveau presque dialectal au contraire, nous avons déjà mis au point un émulateur du changement phonologique sur ordinateur (Lowe et Mazaudon, 1994) soutenant une grande régularité. Nous y chercherons les limites des régularités.

3.2.3 Fennique : segment le plus occidental du groupe finno-ougrien, comprenant une dizaine de langues, où on observe clairement des dérives morphologiques menant du type agglutinant (vepse, finnois) au type fusionnel (estonien, live surtout), sur une profondeur temporelle de trois millénaires.

3.2.4 Finno-Ougrien : cet observatoire est intéressant dans la perspective choisie pour trois raisons : 1) c'est, avec les langues indo-européennes, le domaine premier d'expérimentation de la méthode comparatiste aux 19e et 20è siècle, donc aux archives philologiques les plus fournies ; 2) ce domaine forme un "archipel" continental d'une vingtaine de langues dispersées dans l'Eurasie de la Volga à la Sibérie occidentale en passant par l'Oural, en contact avec les langues indo-européennes et altaïques. En dépit d'un temps de séparation relativement bref (5000 ans), la différenciation y est très forte ; 3) ce groupe a souvent justifié des regroupements hâtifs qui ont la vie dure, tels que "ouralo-altaïque" sur la base d'indices aussi discutables que l'harmonie vocalique ou la structure agglutinante.

3.2.5 Maya (en collaboration éventuelle avec Gilles Polian, Paris III) : ce groupe géographiquement très compact (hormis le huastèque) en Amérique centrale a été relativement bien étudié dans sa fragmentation et ses caractéristique typologiques (travaux de Nora England, Otto Schuman, etc.). En outre, bien qu'il s'agisse d'un groupe linguistique distinct d'autres groupes mesoaméricains, son degré de différenciation rappelle celui d'une famille de langues, avec de fortes continuités dans le lexique et la morphologie verbo-nominale.

3.2.6. Le domaine berbère : ce projet rejoint l'action proposée sur la différentiation génétique et linguistique dans l'aire berbère par J.M. Dugoujon et collaborateurs (N. Bouzekri, R. Ward, Beraud-Colomb etc…) à laquelle participe également Le responsable de ce projet (Pierre Darlu)

Références

ATTENBOROUGH, R., ed., *Papua New Guinea: The human microcosm*, OUP, 1993

BANDELT, HJ, Forster P., Sykes, BC, Richards MB. Mitochondrial protraits of human populations using median networks. *Genetics*, 141 :2, 743-53

BARTHELEMY J.P. et GUENOCHE A., 1991. *Trees and Proximity Representations*, J. Wiley, London, 1991, 238 p

BAXTER William H. et A. MANASTER RAMER, 1996, CR de D. Ringe (1991), *Diachronica* 13/2, p. 371-384.

BENEDICT Paul K., 1973, Tibeto-Burman tones with a note on teleo-reconstruction, *Acta Orientalia* 35, p. 127-138.

BERTORELLE G., EXCOFFIER L., 1998. Inferring admixture proportions from molecular data. *Molecular Biology and Evolution*, 15(10) :1298-1311

BUCK, Carl Darling, 1949, *A Dictionary of Selected Synonyms in the principal Indo-European Languages*, The University of Chicago press.

CAVALLI-SFORZA, L., 1994, *History and Geography of Human Genes*, Princeton.

CAVALLI-SFORA L.L., PIAZZA A., 1975. Analysis of evolution : Evolutionary rates, independence and treeness. *Theoretical Population Biology* 8(2) :127-165.

CHEN J., SOKAL, R.R., Ruhlen M., 1995. Worldwide analysis of genetic and linguistic relationships of human populations. *Human Biology*, 67(4) :592-612

DIXON, R.M.W., 1997, *The Rise and Fall of Languages*, CUP.

FARRIS J.S., KÄLLERSJÖ M., KLUGE A.G., BULT C., 1995. Testing the significance of incongruence. *Cladistics*, 10 :315-319.

GREENBERG Joseph, 1987, *Language in the Americas*, Stanford, Stanford University Press.

HARVEY P.H., PAGEL, M.D., 1991. *The comprative method in evolutionary biology*. Oxford University Press, Oxford, NY

HILL, A. et S. Serjeantson, 1989, The colonization of the Pacific: a genetic trail. Oxford: Clarendon Press.

HOENIGSWALD H.M., WIENER L.F., 1987. *Biological Metaphor and Cladistic Classification. An Interdisciplinary Perspective*. University of Pennsylvania Press

HURFORD, James, et al. (eds.). 1998. *Approaches to the Evolution of Language. Social and cognitive bases.* Cambridge: CUP.

KIRK R., ed., 1989, *Out of Asia.*, Canberra, ANU.

LATHROP G.M., 1982. Evolutionary tree and admixture : phylogenetic inference when ppulations are hyubridized. *Annals of Human Genetics*, 46 :245-55

NETTLE D., 1999. *Linguistic diversity*. Oxford University Press, Oxford, NY

NICHOLS J. 1992. *Linguistic Diversity in Space and Time*. Chicago: University of Chicago Press

NICHOLS J. 1997. Sprung from two common sources: Sahul as a linguistic area. In *Archaeology and Linguistics: Global Perspectives on Ancient Australia*, ed. P McConvell and N Evans, pp. 135-68. Melbourne: Oxford

PAGE, R.D.M., 1990. Temporal congruence and cladistics analysis of biogeography and cospeciation. *Systematic Zoology*, 39 :205-26

PRITCHARD 2000

RINGE Donald, 1991, On calculating the factor chance in language comparison, *Transactions of the American Philosophical Society*, p.1-110.

- 1992, On calculating the factor of chance in language comparison. *Transaction of the American Philosophical Society*, 82 :1-110.
- 1996, The mathematics of 'Amerind', *Diachronica* 13/1, p. 135-154.

ROBINSON D.F., FOULD L.R., 1981. Comparison of phylogenetic trees. *Mathematical Biosciences*, 53 :131-147

RUHLEN M., 1975. *Atlas of the languages of the world* (impression d'auteur, 160p)
RUHLEN, Merritt, 1994 (trad.1997), *L'Origine des Langues*, Débats, Belin
SARGEANTSEN, Susan, ed., 1995, *The Colonization of the Pacific*, Oxford, OUP.
SARGEANTSEN, Susan, R.Kirk and P. Booth, Linguistic and genetic differentiation in New Guinea. Journal of Human Evolution 12.77-92.
SOKAL R.R., 1988. Genetic, geographic, and linguistic distance in Europe. *Proc. Natl. Acad. Sci.* USA 85 :1722-26
WATKINS Calvert, 1990, Etymologies, equations and comparanda: types and values, and criteria for judgment, in : Baldi (ed.), *Linguistic change and reconstruction methodology*, Trends in Linguistics, Studies and Monographs 45, New York, Mouton de Gruyter.

Bibliographie personnelle des collaborateurs relative au sujet

BESSIERE, P., TALBI, E.G., AHUACTZIN, J.M., MAZER, E. (1995) Cooperating parallel genetic algorithms. In : *Handbook of Genetic Algorithms applications*, CRC Press.
BOË, L.-J., SCHWARTZ, J.-L. (1997) L'émergence des structures phonologiques à la lumière des relations production-perception. In : *Perception audtive et compréhension du langage. …Etat initial, Ètat stable et pathologie*, pp. 49-69, sous la dir. de J. Lambert et J.-L. Nespoulous, Solal, Marseille.
DARLU P., TASSY P., 1993. Reconstruction phylogénétique. Concepts et méthodes. Masson, 245pages.
DARLU P., 1988. Are parsimony and Compatibility methods relevant to infer language evolution ? Communication International Meeting on Language change and biological evolution, Turin, 23-25 May.
DARLU P., RUHLEN M., CAVALLI-SFORZA L.L. 1987 A taxonomic analysis of linguistic families. Symposium on Language transmission and changes, Stanford, (W.S.Y. Wang Ed. , still unpublished, 36 pages ronéotées)
DARLU P. 1994. Le poids des hypothèses dans la reconstruction phylogénétique. *Biosystema*, 11 :29-42
FOLEY, W., 1986, *The Papuan languages of New Guinea*, Cambridge University Press.
FOLEY, W., sous presse, The languages of New Guinea, *Annual Review of Anthropology*, c 40 p., 2000
GUÉNOCHE A., PRÉA P..1998- Counting and selecting at random phylogenetic topologies to compare reconstruction methods, Proceedings of the Conference of the International Federation of the Classifications Societies, IFCS'98, Short papers volume, pp. 242-245
GUENOCHE A. , GARRETTA. H., 2000 - Quelle confiance accorder à une représentation arborée ?, in Actes des Journées Biologie, Informatique Mathématiques,Agro-Montpellier, 2000, pp. 181-188.
JACOBSON, M, 1999, "Automatisation du découpage syllabique par l'utilisation d'un canon", Actes du colloque SyllabeS (Nantes, mars 1999).
LECOINTRE G., RACHDI L., DARLU P., DENAMUR E., 1998. E. Coli molecular phylogeny using the Incongruence length difference test. Molecular Biology and Evolution 15(12) :1685-1695.
LEONARD, J.L., 1999 : "Aspects de la ptosigenèse dans les langues finno-ougriennes" in Histoire, Épistémologie, Langage, 2000 n° spécial "Formation de la Syntaxe".
LEONARD, J.L., 1999 : "Aires dialectales, naturalité et options structurales étagées : l'exemple du fennique", colloque Lacito, CNRS, organisé par Jocelyne Fernandez-Vest, avril 99, à paraître.
LEONARD, J.L., 2000 : "L'observatoire dialectal d'oïl, un domaine à explorer pour les nouvelles phonologies", colloque de phonologie du GDR dirigé par Bernard Laks, Bordeaux, juin 2000.
LOWE John B. and Martine MAZAUDON, 1994, The Reconstruction Engine: A computer implementation of the comparative method. *Special Issue on Computational Phonology, Computational Linguistics* 20:3.
MARSICO E., 1999 "What can a database of proto-languages tell us about the last 10,000 years of sound changes ? "Poster, XIVth Int.ernational Congress of Phonetic Sciences, San Francisco, Aout 1999 Publié dans les actes du colloque
MARSICO, E , HOMBERT J.M., 2000. "(Re) Evaluating the implicit assumptions behind historical linguistics methods" Communication, Conference "Long-Range Linguistic Comparison : Prospects On The Eve Of The Third Millennium" Moscou, 29 Mai - 3 Juin 2000. Publie dans les actes
MARSICO E., COUPE C., et PELLEGRINO F., 2000. "Evaluating the influence of language contact on lexical changes" Poster, The Evolution of Language Conference Paris, Avril 2000 Publie dans les actes

MATISOFF, James A., 1978, *Variational Semantics in Tibeto-Burman*, Occasional Papers of the Wolfenden Society on Tibeto-Burman Linguistics, Philadelphie, Institute for the Study of Human Issues.
MATISOFF, James A., 1985, God and the Sino-Tibetan copula, *J. of Asian and African Studies* 29:1-81. Tokyo.
MATISOFF, James A, 1994, How 'dull' can you get? 'buttock' and 'heel' in Sino-Tibetan, *LTBA* 17/2: 137-151.
MAZAUDON Martine et John B. LOWE, (dec 1993, sous presse), Regularity and exceptions in sound change, in Dominicy et Demolin, eds, *Investigations in Sound Change*, Oxford University Press.
MAZAUDON M. et J. B. LOWE, 1991, Du bon usage de l'informatique en linguistique historique, *BSLP* 86/1, p. 49-87.
PATRICK, Jon, 1977a, "A complexity measure for diachronic Chinese phonology", *Proceedings of the SIGPHON97 workshop on Computational Linguistics at the ACL'97/EACL'97*, Madrid, Spain.
PATRICK, Jon, 1997b, "Linguistic similarity measures using the Minimum Message Length principle", *Archaeology and Language I: Theoretical and Methodological Orientations*, (eds.) R. Blench & M. Spriggs. London: Routledge. pp 260-277.
REBUSCHI, Georges, 1994. Linguistique historique et comparative et nouvelle grammaire comparée: réflexions épistémologiques; *in* J.B. ORPUSTAN (éd.), *La langue basque parmi les autres; Influences et comparaisons*, Baïgorry, Izpegi, 149-185.
SCHWARTZ, J.L., BOË, L.J., VALLÈE, N., ABRY , C. (1997) Major trends in vowel system inventories. *Journal of Phonetics*, 25, 233-254.
SCHWARTZ, J.L., BOË, L.J., VALLÈE, N., ABRY , C. (1997) The dispersion-focalization theory of vowel systems. *Journal of Phonetics*, 25, 255-286.
VALLEE N., BOË L.J., STEFANUTO M. (1999). Typologies phonologiques et tendances universelles. Approche substantialiste. Linx, 31-54. Numéro spécial de 1999