Guillaume Segerer **LLACAN - UMR8135**
Dossier de demande d'intégration dans le corps des CR - avril 2008

Projet de recherche

Comparaison des langues atlantiques génétique - typologie - universaux

Projet de recherche

Comparaison des langues atlantiques
génétique - typologie - universaux

Présentation

Mon projet de recherche prend comme champ d'investigation le domaine particulier des langues "atlantiques" (groupe linguistique de l'ouest de l'Afrique), pour apporter une contribution dans trois domaines :

(1) l'histoire des langues d'Afrique en général et d'Afrique de l'Ouest en particulier,
(2) la théorie et les méthodes en linguistique historique,
(3) la typologie générale et la recherche des universaux linguistiques.

Il s'appuie sur une connaissance solide des structures des langues atlantiques, une expérience de la linguistique de terrain – nécessaire pour combler les lacunes de la documentation, notamment sur les langues vouées à une extinction rapide – mais aussi, et c'est l'un des aspects originaux de la démarche adoptée, sur des compétences reconnues dans le domaine des bases de données et de la programmation à l'usage du web.

L'originalité de ce projet repose en effet sur l'utilisation intensive de l'informatique, non seulement pour le stockage, la diffusion et la visualisation des données, mais aussi, et surtout, comme instrument d'aide à la décision.

Après avoir évoqué l'état de confusion qui prévaut dans le domaine du comparatisme atlantique, je présenterai mon projet proprement dit. Prenant comme point de départ la méthode comparative classique, il vise à augmenter son efficacité par l'utilisation de techniques d'investigation innovantes.

1. Langues atlantiques : un groupe linguistique problématique

a) Etat de la classification

La branche *atlantique*[1] des langues Niger-Congo forme un ensemble d'environ 60 langues. Celles-ci occupent une bande côtière relativement étroite entre le nord du Sénégal et le nord-ouest du Liberia (*cf.* carte ci-dessous), où elles sont souvent en contact avec des langues du groupe Mande.

Fig. 1. Extension des langues atlantiques[2]
(d'après les cartes disponibles à www.ethnologue.com)

[1] Dans ce qui suit, les noms de langue sont en minuscule et les noms de groupes de langues prennent une majuscule. Deux exceptions cependant : le deuxième élément du nom d'une langue prend une majuscule s'il s'agit d'un nom propre (ex. : joola Kasa) ; le nom d'un groupe utilisé comme adjectif est en minuscules s'il est accordé (ex. langues atlantiques / parlers Joola).

[2] Cette carte concerne la zone proprement atlantique. Le wolof est également parlé en Mauritanie ; diverses variétés de peul sont pratiquées dans la quasi-totalité de la bande sahélienne : Mali, Burkina Faso, Niger, Nigeria, Cameroun, Tchad, République Centrafricaine, Bénin, Togo, Ghana, etc.

Cet ensemble, dont les contours ont été esquissés dès 1854 par S. Koelle (*Polyglotta Africana*), précisés par Westermann (1928) avant de se fixer avec Greenberg (1963), demeure une épine dans le pied des (rares) comparatistes africanistes. En effet, son unité génétique n'a toujours pas été démontrée et n'est postulée que sur la base de critères géographiques et typologiques.
La structure interne du groupe, établie par D. Sapir (1971) et reproduite depuis avec parfois des modifications mineures[3], reconnaît trois zones : Nord, Sud, et Bijogo[4].

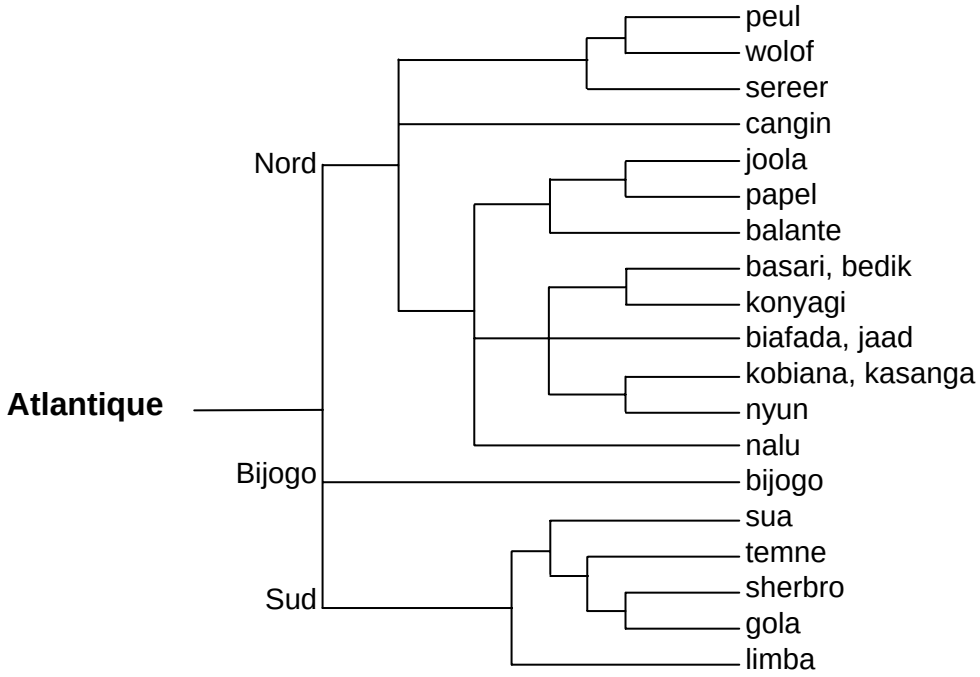

Fig. 2. Classification généalogiques des langues atlantiques
(version de Wilson 1989 reprise dans Blench & Williamson 2004)

b) Le paradoxe atlantique

Cette structuration interne repose uniquement sur des calculs lexicostatistiques. Elle n'est pas remise en cause pour la raison suivante : comme il est encore impossible d'apporter la preuve de la validité génétique de la famille atlantique, il est difficile, à plus forte raison, de revenir sur le détail de son organisation interne.
En effet, de D. Dalby (1965) à T. Childs (2001) en passant par W.A.A. Wilson (1989, 2007) ou J.L. Doneux (1975, 1991), les spécialistes s'accordent à considérer l'unité de cette famille comme extrêmement douteuse. Il est vrai que les mesures lexicostatistiques ne sont guère encourageantes : le taux de ressemblances entre deux langues dites atlantiques est souvent inférieur à 10%, et même à l'intérieur d'un sous-ensemble réputé non problématique, comme le joola, on trouve des valeurs inférieurs à 20% (entre le *bayot* et n'importe quel autre parler joola). Cependant, on ne peut que constater, en parcourant la littérature des

[3] Sans entrer dans les détails, il n'est pas inutile de signaler que les raffinements de la classification de 1971 ne sont jamais argumentés.
[4] Cette langue, dont j'ai proposé une description (Segerer 2002a), constitue un isolat.

quarante dernières années (les sources antérieures disposaient de très peu de données), que bien peu a été fait pour résoudre ce paradoxe. Les quelques tentatives pour imposer une réorganisation généalogique, comme par exemple Dalby (1965) qui suggère d'écarter le groupe Mel, n'ont pas été prises au sérieux par la communauté scientifique qui continue de citer Greenberg, oubliant également Doneux (1975), responsable de la simplification en *Atlantique* de l'ancienne étiquette *Ouest-Atlantique*.

Dans l'usage, certains petits groupes de langues ont été reconnus comme présentant une plus grande cohérence interne : ces ensembles occupent dans la figure ci-dessus des positions diverses :

– l'appellation *Cangin* renvoie en fait à cinq langues proches et néanmoins distinctes : le *lehar*, le *ndut*, le *noon*, le *palor* et le *safen* ;

– le groupe *Bak* est représenté ci-dessus par 3 éléments : *joola*, *papel* (un des parlers *manjaku*) et *balante*, le *bayot* étant implicitement inclus dans le groupe Joola ;

– les lignes *basari, bedik* et *konyagi* forment avec le *bapen* (non cité) l'ensemble *Tenda* ;

– le sous-groupe *Sénégambien*, appellation aujourd'hui tombée en désuétude, comprenait le *peul*, le *sereer*, le *wolof* et les langues *Cangin* ;

– au sud, on nomme *Mel* l'ensemble comprenant les langues *kisi* (non cité ci-dessus)*, sherbro*, *temne, baga* et *gola*, c'est-à-dire presque toute la zone.

Derrière les langues les plus méconnues peuvent se dissimuler des ensembles dialectaux parfois complexes. Ainsi le *nalu*, sur lequel j'ai pu faire une très courte enquête en 1998, semble-t-il pouvoir être divisé en au moins trois parlers distincts. De même certains des dialectes du *bijogo* pourraient sans doute être considérés comme des langues distinctes, quoique proches.

Les langues du groupe Nord sont les plus nombreuses et les mieux documentées. On relève cependant de grandes disparités dans la nature et la qualité des données disponibles. Ainsi par exemple, alors que le *peul* et le *wolof* font l'objet de descriptions depuis plus d'un siècle et possèdent déjà une solide tradition linguistique, on ne dispose pour le *biafada* ou le *nalu* que de courtes listes de mots. Dans la zone sud, la situation est comparable. Seul le *kisi* a fait l'objet d'une description récente (Childs 1995). Pour d'autres langues comme le *gola* ou le *temne*, les références sont soit anciennes et parfois introuvables (Westermann 1921, Schlenker 1864, Scott 1965), soit limitées à des domaines particuliers (par ex. Koroma 1994). Le *limba* n'est connu que par un dictionnaire de 1922, les parlers *baga* sont très mal documentés, et pour le *sua* il n'existe que trois listes de mots non publiées : les deux premières sont dues à W.A.A. Wilson (com. pers.) et J. L. Doneux (s.d.) ; j'ai personnellement recueilli la troisième à Bissau en 1998.

c) Des difficultés sérieuses

Les langues atlantiques sont très diverses mais en même temps elles affichent, lorsqu'on les compare aux groupes voisins (Mande et Kru) une unité assez convaincante basée notamment sur la présence de classes nominales, une dérivation verbale productive et la prépondérance du type CVC dans les racines lexicales.

Cependant, les modalités concrètes de ces traits typologiques peuvent présenter des différences extrêmes : c'est ainsi que, dans les lexèmes nominaux, les marques de classes nominales sont parfois quasi absentes (*wolof*), préfixées (*bijogo*, *joola*), suffixées (*peul*, *kisi*), préfixées *et* suffixées (*sereer*). Des évolutions divergentes compliquent encore la situation : ainsi, les marques de classes nominales, très probablement de forme CV à l'origine, ont perdu leur consonne dans le groupe Tenda et leur voyelle en *wolof*. Les marques de classes actuelles des langues Tenda ne présentent donc plus rien de commun avec celles du *wolof*.

Ces phénomènes d'érosion touchent également les racines lexicales. Ainsi observe-t-on parfois, dans les langues où les classes nominales sont marquées par des préfixes, une tendance des unités lexicales à s'éroder par la droite, les préfixes étant alors réinterprétés comme des éléments de racines et se voyant affecter de nouveaux préfixes. D'une manière générale, les interactions entre les racines lexicales 'originales' et les affixes ont pour effet de masquer les ressemblances (voir notamment Pozdniakov & Ferry 2001 pour des exemples détaillés).

L'existence de telles disparités rend la comparaison extrêmement problématique et explique sans doute pourquoi personne jusqu'à maintenant n'a pu apporter la preuve de l'unité génétique de cette ensemble. Ces variations sont aussi en partie responsables de la grande diversité lexicale de ces langues.

d) Des solutions à portée de la main

Aujourd'hui, le comparatisme atlantique peut enfin envisager des progrès réels, pour trois raisons :

– Les principales difficultés morphologiques sont identifiées. La partie proprement atlantique de ce projet vise entre autres à l'étude systématique des processus morphologiques du type de ceux évoqués ci-dessus, en vue de dégager les radicaux pour la comparaison lexicale mais également pour mieux comprendre les types d'évolution morphologique de ces langues.

– La masse des données lexicales disponibles n'a jamais été aussi importante. J'ai réuni, au cours des dix dernières années, un corpus informatisé de plus de 60000 entrées lexicales, représentant l'ensemble des 19 branches de la figure 2 ci-dessus (p. 4). Ce corpus contient non seulement près des trois quarts des données publiées à ce jour, mais également une quantité importante de données inédites : enquêtes personnelles (*bijogo*, *manjaku*, *nalu*, *sua*), littérature grise, manuscrits originaux (provenant notamment du fonds Doneux, conservé au LLACAN). J'ai évidemment l'intention de continuer à enrichir cette ressource unique, à la fois par la saisie des

sources existantes et par des enquêtes de terrain pour les langues les moins documentées.
– Enfin, des outils informatiques spécifiques sont maintenant disponibles pour exploiter ces bases de données et les premiers résultats sont là : deux branches (tenda et joola) sont en ce moment systématiquement explorées et des centaines de séries comparatives ont déjà été établies. Aujourd'hui, les programmes que j'ai mis au point fournissent une assistance précieuse à la recherche de cognats, et permettent une gestion rigoureuse des séries comparatives ainsi que de hypothèses de reconstruction. Je projette d'augmenter l'efficacité de ces outils et d'en faire des instruments complets d'aide à la décision.

Les techniques employées pour parvenir à ces résultats nourrissent inévitablement une réflexion théorique sur la méthode comparative. L'étude des conséquences méthodologiques et théoriques des méthodes modernes de traitement des données constitue le deuxième volet de ce projet de recherche.

2. Nouvelles méthodes pour le comparatisme

Les principes de la méthode comparative classique, s'ils demeurent nécessaires à toute entreprise de reconstruction linguistique, s'avèrent parfois insuffisants pour traiter certaines situations complexes (cf. notamment Durie & Ross (eds) 1996). Ce volet de mon projet de recherche consiste précisément à proposer des méthodes originales pour aborder la complexité. Cette démarche se développe pour l'instant dans deux directions : la statistique phonologique et les représentations phylogénétiques.

a) Statistique phonologique

L'utilisation des statistiques dans un cadre comparatiste n'est pas en soi une démarche révolutionnaire mais s'avère encore marginale[5]. Pourtant, les techniques utilisées sont simples et très efficaces. Elles permettent par exemple de formuler des prédictions quant aux types de correspondances régulières à rechercher, et ceci avant même tout travail de comparaison.
En voici un exemple. Les deux tableaux ci-dessous montrent les fréquences relatives des différents types de combinaisons <consonne-voyelle> dans deux parlers proches du groupe Joola, le Kasa et le Fogny. Les consonnes comme les voyelles ont d'abord été regroupées en classes d'affinité en fonction de leur point d'articulation dominant, puis les combinaisons des différentes classe de consonnes avec les différentes classes de voyelles ont été comptées[6]. On voit immédiatement

[5] Voir Pozdniakov (1991) pour une bonne illustration de l'intérêt de cette approche.
[6] Dans les tableaux, les classes de phonèmes sont représentées par des lettres majuscules : P= consonnes labiales, T= consonnes dentales/alvéolaires, C= consonnes palatales, K= consonnes vélaires, A= voyelles centrales, I= voyelles antérieures, U= voyelles postérieures. Les chiffres signalent l'écart (en %) entre le nombre de combinaisons observées et le nombre de combinaisons attendues. Les couleurs permettent de repérer les écarts significatifs : couleur claire = écart entre 15% et 30%, couleur foncée = écart supérieur à 30% ; bleu = écart négatif, orange = écart positif.

que les deux langues présentent des situations très différentes, en particulier quant au traitement des séquences CV où C est une vélaire et V une voyelle antérieure.

Fogny	I	A	U
P	-8	8	-1
K	-94	23	55
T	20	-3	-13
C	57	-25	-23

joola Fogny (3498 séquences CV)

Kasa	I	A	U
P	-14	6	9
K	-2	-5	6
T	10	-3	-8
C	-3	8	-4

joola Kasa (2190 séquences CV)

Fig. 3. Séquences CV dans deux parlers Joola

En Fogny, il manque 94% de séquences KI par rapport à la norme calculée sur la base des fréquences réelles de K initial et I final dans les séquences CV[7]. En outre, on observe un excès important des séquences CI (consonne palatale-voyelle antérieure) et KU (consonne vélaire - voyelle postérieure). Cette distribution n'est pas observée en Kasa, où tous les écarts par rapport aux valeurs attendues sont à l'intérieur de limites raisonnables.

Ces deux tableaux permettent de prévoir avec un degré de certitude important qu'aux séquences KI du Kasa vont correspondre des séquences CI ou KU en Fogny. En outre, on peut également poser que l'évolution s'est plutôt faite à partir de *KI (>CI ou >KU) et pas l'inverse, ce qui correspond à une assimilation.
L'examen des lexiques[8] de ces langues montre qu'il existe effectivement une correspondance régulière entre Kasa KI et Fogny CI :

	Kasa	Fogny
veine	ka-kil	ka-cil
déchirer	-gis	-jis
mourir	-kèt	-cet
insulter	-gèl	-jel

Cet exemple est volontairement trivial. Les correspondances impliquées sont faciles à trouver, et les deux langues sont très proches. Mais il montre qu'il est possible d'extraire des informations sur une protolangue **sans comparer directement** les mots des langues actuelles. Quelles sont les limites d'une telle approche ? Quelles autres informations peuvent-elles être obtenues de cette façon ? Les statistiques peuvent-elles être utiles dans d'autres domaines que la phonologie ? C'est à ce type de question que ce volet de mon projet est consacré.

b) Représentations phylogénétiques

Traditionnellement, les relations généalogiques sont représentées sous forme d'arbres plus ou moins complexes. Ces représentations peuvent être utiles mais leur inconvénient principal est qu'elles suggèrent des processus d'évolution souvent mécaniques et sans nuances. En fait, elles véhiculent des hypothèses très

[7] Sur 3498 séquences CV, 1033 (29,5%) ont un K initial et 586 (16,8%) ont un I final. En l'absence de toute corrélation, le nombre de séquencse KI observées devrait être : 3498 x 29,5% x 16,8% = 173. Or, on n'en compte que 10, soit un déficit de 94%.
[8] Pour le Kasa : Wintz 1909 ; pour le Fogny : Sapir 1970.

fortes sur l'histoire des langues. Prenons un exemple avec les parlers Joola, pour lesquels il existe des données lexicostatistiques concernant 31 parlers (Carlton & Rand 1993). Traitées par le logiciel LEXISTAT, ces données permettent d'obtenir un arbre 'classique' :

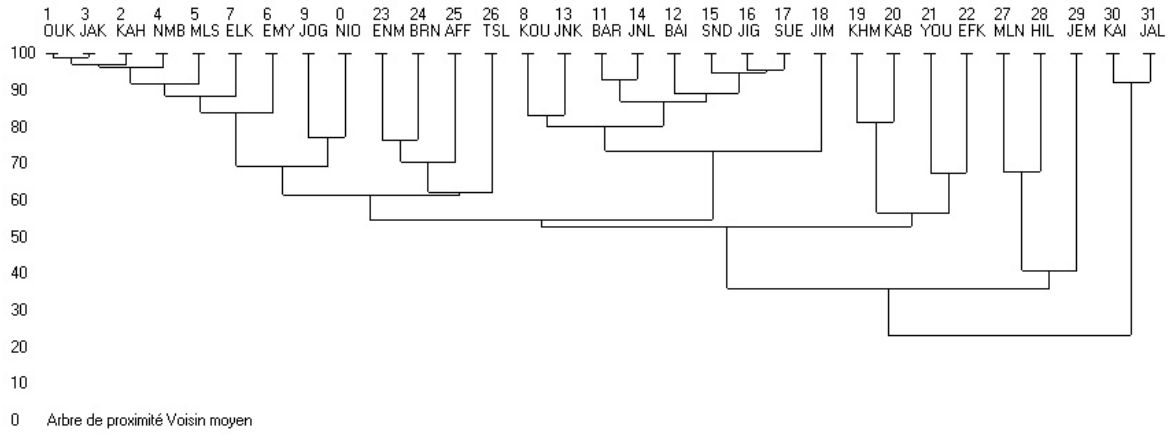

Fig. 4. Les parlers Joola (méthode LEXISTAT)
(d'après Carlton & Rand 1993)

Parmi les informations qu'apporte cet arbre, on note les fait suivants :

– Il existait probablement une protolangue (tout est relié)

– Celle-ci s'est divisée très tôt, isolant un groupe formé des parlers 30 et 31 (Kaïlou et Dialang).

– Chaque étape de l'évolution est une division en deux.

– Toutes les langues sont également distantes de la racine, et l'échelle de pourcentages de ressemblances à gauche peut être interprétée comme une échelle chronologique.

Toutefois, on aimerait pouvoir disposer, à partir des mêmes données, de représentations illustrant des conceptions différentes de l'évolution. Dans ce domaine, les linguistes accusent un retard considérable sur les biologistes, par exemple. Ces derniers développent depuis longtemps des outils permettant de traiter d'importantes masses de données et de les représenter, et ceci non seulement en préservant leur complexité, mais aussi en permettant diverses interprétations des mêmes faits.

En fait, on assiste actuellement à un renouveau des pratiques et quelques tentatives sont faites pour appliquer au champ linguistique les techniques développées initialement pour la biologie. Pour le domaine africain, ce sont les langues Bantu qui ont les premières fait l'objet de tels traitements (Russel & Gray 2006). Les résultats obtenus ont permis de formuler de nouvelles hypothèses sur l'histoire de ces langues. Cete partie de mon projet consiste non seulement à appliquer ces techniques aux langues atlantiques, mais aussi à mener une réflexion plus théorique sur les valeurs méthodologiques des différents types de représentation de l'évolution.

Pour illustrer l'apport attendu de ces méthodes, je prendrai le cas des parlers Joola. J'ai cette fois introduit les données lexicostatistiques ayant servi à produire

l'arbre de la fig. 4 ci-dessus dans le programme SPLITSTREE[9], moyennant un formatage adapté[10]. Cet outil permet de construire différents types d'arbres, qui fournissent chacun un type particulier d'information. En voici un :

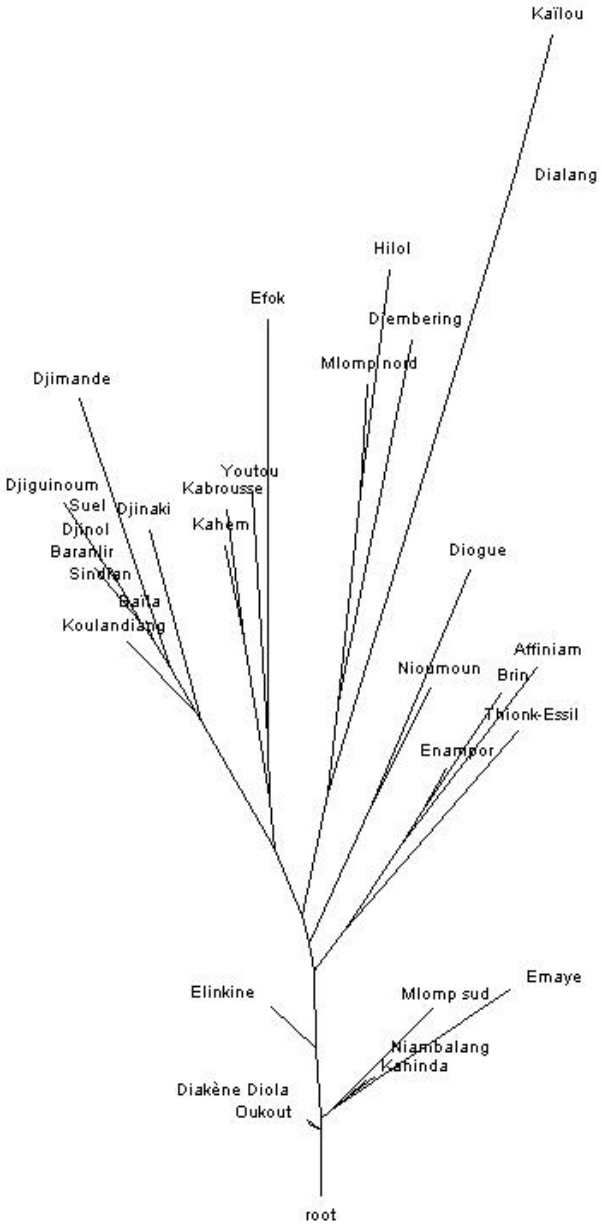

Fig. 5. Les parlers Joola (méthode SPLITSTREE 1)
(d'après Carlton & Rand 1993)

Ici, et bien que les données soient les mêmes, les informations obtenues sont différentes de celles fournies par l'arbre 'classique' (fig. 4). Si la protolangue est toujours postulée, la première division est toute autre (les parlers Kaïlou et Dialang sont maintenant parmi les derniers à diverger). En outre, à chaque étape de l'évolution, on n'observe plus une division en deux, mais plutôt la pousse d'une nouvelle branche. En outre, cette représentation suggère que certaine langues

[9] http://www.splitstree.org/
[10] Ce programme est initialement conçu pour traiter des séquences de nucléotides.

sont plus proches que d'autres de la 'racine', c'est-à-dire plus conservatrices. Considérons maintenant une deuxième représentation obtenue avec SPLITSTREE :

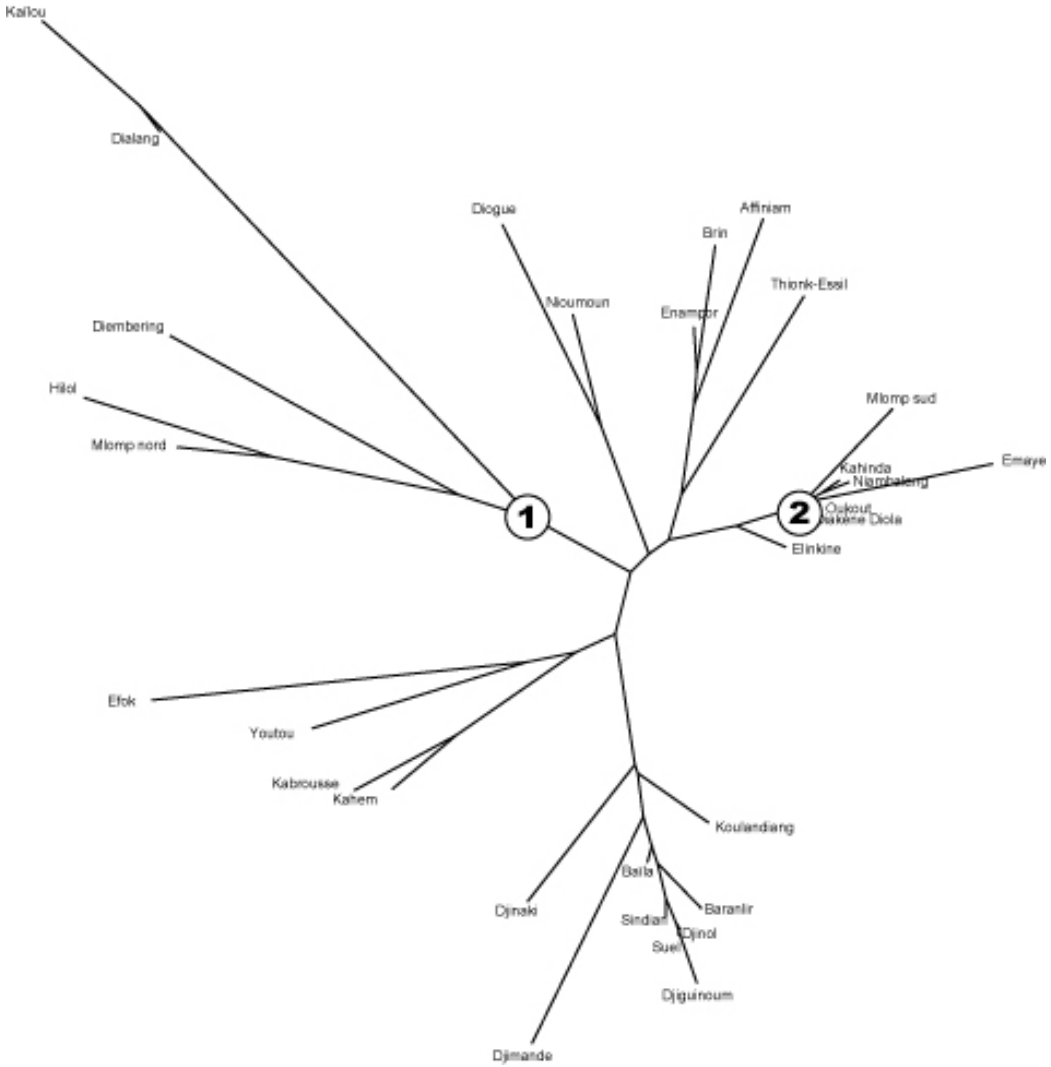

Fig. 6. Les parlers Joola (méthode SPLITSTREE 2)
(d'après Carlton & Rand 1993)

Ici la protolangue est seulement implicite (puisque tout est relié), mais sa position n'est pas précisée. Les pastilles numérotées 1 et 2 signalent les positions de la racine (la protolangue) respectivement pour la méthode LEXISTAT et la méthode SPLITSTREE 1. De plus, les divisions ne sont plus successives, comme avec les deux précédents arbres, mais peuvent être comprises comme s'étant produites plus ou moins simultanément. Enfin, la répartition spatiale des langues sur la figure n'est que faiblement explicite quant au détail de la chronologie.

On le voit, le choix d'une méthode de représentation de l'évolution n'est pas sans conséquence sur les hypothèses concernant l'évolution elle-même. Une partie de ce projet est l'étude des possibilités offertes par les outils existant et à venir, mais il s'agit aussi de comprendre les logiques qui sous-tendent l'élaboration des divers types de représentation.

3. Typologie et universaux

Le dernier volet de mon programme de recherche est consacré à la typologie des langues africaines et à la recherche d'universaux linguistiques. L'intérêt pour la typologie découle de son importance pour la comparaison, qui ne peut pas être envisagée du seul point de vue lexical. La recherche des universaux, en revanche, s'inscrit dans un questionnement sur la portée des phénomènes observés.

a) Terrain et description

J'envisage de poursuivre la documentation des langues atlantiques, car si l'on dispose maintenant de nombreuses ressources lexicales, il en va tout autrement pour les descriptions grammaticales. Sur les quelques 60 langues atlantiques, une dizaine seulement est dotée d'une description fiable. Dans un premier temps, je compte me consacrer à un parler Joola de Casamance (une mission est prévue dès février 2008), mais j'ai également l'intention de travailler sur des langues méconnues de Guinée Bissau, comme le *sua* ou le *nalu*.

Par ailleurs, ma participation à plusieurs projets collectifs internationaux (marques personnelles, adjectifs et qualification) m'a conduit à m'intéresser aux autres familles de langues en Afrique, et j'entends bien poursuivre ce type d'investigation.

b) Statistique

En testant des hypothèses de travail sur les langues atlantiques, mon collègue Konstantin Pozdniakov et moi-même avons observé des restrictions dans les possibilités combinatoires des consonnes au sein des racines lexicales. Croyant avoir enfin trouvé un critère pour affirmer la parenté de ces langues, nous avons effectué des mesures sur d'autres langues, sans jamais rencontrer de contre-exemple. Après de nombreuses autres mesures et vérifications, nous avons pu formuler une généralisation : les langues naturelles présentent une tendance marquée à éviter les combinaisons de type CVC dans lesquelles les deux consonnes sont de même point d'articulation[11].

Ce phénomène était déjà bien connu pour les langues sémitiques en raison de la nature 'gabaritique' des racines lexicales de ces langues. La nouveauté est de l'avoir mis en évidence pour de nombreuses autres langues en utilisant des méthodes statistiques simples.

Les tableaux ci-dessous[12] montrent que cette tendance n'est pas moins marquée pour le français, par exemple, que pour l'arabe (résultats inédits) :

arabe	P	K	T	C
P	-43	16	4	11
K	20	-52	12	9
T	19	21	-15	-1
C	20	-10	10	-34

81532 mots et 179378 séquences CVC)

français	P	K	T	C
P	-58	15	12	2
K	36	-35	4	16
T	17	10	-11	-4
C	-6	2	7	-35

(43260 lemmes et 81382 séq. CVC)

[11] Cette découverte a été récemment publiée dans la revue *Linguistic Typology* (Pozdniakov & Segerer 2007).
[12] Les conventions graphiques sont les mêmes que pour les tableaux de la p. 8.

Ces résultats sont inattendus. Quel locuteur du français, en effet, a conscience que sa langue évite les séquences -PVP- (consonne labiale - Voyelle - consonne labiale) ? N'importe qui peut citer instantanément des dizaines de mots formés sur ce modèle : *pomme*, *pape*, *fève*, *femme*, *baume*, *bombe,* etc. Pourtant ces séquences sont moins nombreuses qu'elles ne le seraient en l'absence de toute corrélation, et ceci dans des proportions tout à fait significatives : la base de données LEXIQUE (www.lexique.org), qui a été utilisée pour élaborer le tableau ci-dessus, contient précisément 1746 de ces séquences. Si deux consonnes labiales pouvaient se combiner librement, on devrait en trouver 4142[13]. Pour les langues atlantiques, des calculs plus détaillés montrent que des phénomènes annexes, comme la longueur vocalique ou la présence d'une frontière morphologique ont une influence directe sur les possibilités combinatoire des consonnes. Les conséquences de ces découvertes sur notre vision de l'histoire des langues seront sans aucun doute très importantes.

c) Sémantique

Une autre partie de ce programme est consacrée à la modélisation du sens et à la recherche de tendances générales dans l'organisation de l'espace sémantique. Dans un premier temps, il s'agit d'appliquer à des langues diverses l'outil exploratoire PROX[14], développé par Bruno Gaume (CNRS, ERSS-IRIT, Toulouse). La version actuelle de PROX permet de visualiser une projection tridimensionnelle de l'espace sémantique associé à un élément lexical du français[15], sur la base des relations synonymiques fournies par sept dictionnaires.

On peut exploiter les mêmes méthodes avec d'autres langues que le français mais aussi avec d'autres types de relations que la synonymie. Pour l'instant, des essais ont été faits avec les relations polysémiques à l'intérieur de mon corpus atlantique. Le graphe obtenu a été comparé à celui du français. Cette démarche soulève un certain nombre de questions théoriques et pratiques qui devront être éclaircies : est-il légitime de comparer un graphe monolingue de synonymes et un graphe multilingue de polysèmes ? L'influence du français sur la structure du graphe atlantique est-elle mesurable, et ne compromet-elle pas l'accès aux structures propres aux langues étudiées ?

Les premiers résultas obtenus indiquent que malgré le recours à une langue pivot (ici le français), le graphe atlantique présente des particularités propres, et les similarités de structure peuvent être attribuées au type d'espace considéré (ici l'espace sémantique). Les deux figures ci-dessous, obtenues avec PROX à partir du graphe atlantique, représentent respectivement le voisinage sémantique direct de la notion *sentir* (c'est-à-dire l'ensemble des termes qui figurent au moins une fois à

[13] Les consonnes labiales (P) sont à l'initiale de 26,8% des séquences CVC, et en finale de 18% de ces séquences. La base LEXIQUE contient 85857 séquences CVC. On doit donc s'attendre à trouver 18% x 26,8% x 85857 = 4142 séquences PVP.
[14] http://erss.irit.fr/prox/
[15] Voir notamment Gaume (2004) pour le détail de son fonctionnement.

côté de *sentir* dans la définition d'un terme de l'une des langues du corpus), et les 50 premiers sens associés (la proxémie-50, soit les 50 sommets du graphe qui sont les plus proches, topologiquement, du sommet *sentir*).

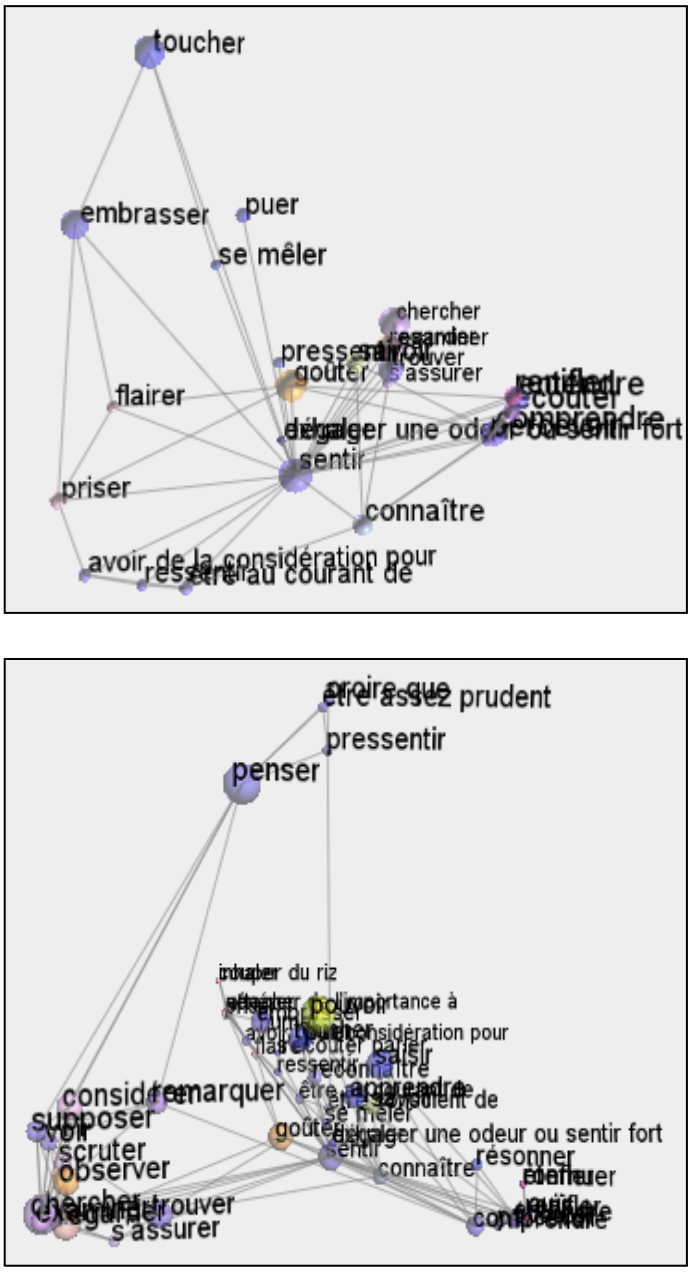

Fig. 7. Voisinage et proxémie de *sentir*

Ce que montrent ces images, c'est par exemple que la notion *penser* (qui n'est pas un voisin de *sentir*, c'est-à-dire qui n'est jamais associée à *sentir* dans une relation polysémique pour les langues atlantiques considérées) fait néanmoins parties des notions les plus proches. Ce type de résultat peut être exploité pour étendre la portée des recherches de cognats, ou pour argumenter *a posteriori* sur la validité de rapprochements sémantiques opérés sur des bases purement formelles.

Cette expérience a fait l'objet d'une communication que Bruno Gaume, Martine Vanhove et moi-même avons présenté au dernier colloque de l'ALT (Paris, sept. 2007).
Un prochain développement de ces travaux est la mise au point, avec l'aide de Bruno Gaume, d'un outil permettant de mesurer la *distance sémantique* entre les mots. Un tel outil devra permettre de formuler des hypothèses, puis des généralisations, sur le caractère universels de certains phénomènes de polysémie, de métaphore, d'évolution du sens.
Dans le même temps, il pourra être utilisé directement pour la recherche comparative. Si l'on considère conjointement la distance sémantique et la distance phonétique, et si l'on se donne la possibilité de faire varier l'influence de l'une ou de l'autre, alors il devient possible de rechercher automatiquement, et avec une grande finesse, des éléments susceptibles d'être apparentés.

4. Des conséquences globales

L'étude détaillée d'un groupe de langues à l'histoire complexe, et surtout les modalités de cette étude, auront des conséquences plus larges. D'une part, il s'agit de confirmer la **validité de la méthode comparative classique** en la dotant de moyens modernes. Il va de soi que les outils développés dans le cadre de ce projet seront utilisables pour d'autres groupes de langues. Il sera également facile de mettre au point une interface de communication avec d'autres bases de données, comme par exemple la base UNIDIA des changements phonétiques[16]. J'ai également l'intention de présenter un projet à l'ANR autour d'une base de données lexicale des langues d'Afrique, projet qui bénéficiera de toute l'expérience accumulée dans le traitement de corpus bilingues.

Cependant, ce projet vise également les objectifs suivants, plus ambitieux :
– **Comprendre** et **modéliser** les processus d'évolution et de divergence dans les langues naturelles. Plus précisément, je cherche à mettre au point de véritables instruments de mesure qui permettront d'évaluer les parts respectives des différents processus dans l'évolution : changements phonétiques réguliers, emprunt, diffusion, sans oublier le poids des contraintes 'universelles' ou locales. Parmi ces dernières, je place non seulement les contraintes phonologiques évoquées ci-dessus, mais aussi toutes les tendances à l'unification ou au contraire à la différenciation qui ont pour cadre les divers sous-systèmes (pronoms, classes nominales, etc.).
– Explorer les **structures non évidentes** dans les langues, que ce soit au niveau de la forme ou du sens. Par 'structures non évidentes', j'entends les traits d'organisation qui échappent non seulement à la conscience des locuteurs, mais aussi aux méthodes d'investigation habituelles. Les résultats obtenus m'encouragent dans cette voie.

[16] http://www.diadm.ish-lyon.cnrs.fr/unidia/index.php.

Dans ces deux axes de recherche, les méthodes utilisées assurent la **reproductibilité** des résultats, mais permettent également de faire des **prédictions**, satisfaisant ainsi aux exigences de la rigueur scientifique. Mon objectif à long terme est de montrer que les systèmes les plus complexes peuvent, dans une large mesure, être compris comme des combinaisons de sous-systèmes plus simples.

5. Ouvrages cités

BARRY, Abdoulaye. 1987. *The Joola languages: subgrouping and reconstruction*. London: School of Oriental and African Studies, University of London.

BLENCH, Roger & Kay WILLIAMSON. 2004 [2000]. 'Niger-Congo', *in* HEINE, B. & D. NURSE (éds), *Les langues africaines*, Paris : Karthala, Agence universitaire de la francophonie, pp. 21-54, traduction G. Segerer.

CARLTON Elizabeth M. & Shanon RAND. 1993. *Enquête sociolinguistique sur les langues diolas de Casamance*. Dakar : SIL, Cahiers de recherche linguistique 2.

CHILDS, G. Tucker. 1995. *A Grammar of Kisi : A Southern Atlantic Language*. Berlin, New-York : Mouton.

CHILDS, G. Tucker. 2001. *What's so Atlantic about Atlantic?* Communication présentée au 31e Colloquium on African Languages and Linguistics, University of Leiden.

CLARKE, M. L. 1922. *Limba-English dictionary*. Freetown : Government Printer.

DALBY, David. 1965. The Mel languages: a reclassification of southern 'West Atlantic'. *African language studies*, 6, pp. 1-17.

DONEUX, Jean-Léonce. 1975. Hypothèses pour la comparative des langues atlantiques. Tervuren, MRAC, *Africana Linguistica VI*, 88, pp. 41-129.

DONEUX, Jean Léonce. 1991. *La place de la langue buy dans le groupe atlantique de la famille kongo-kordofan*. Bruxelles: Université Libre de Bruxelles.

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

GAUME, Bruno. 2004. Balades Aléatoires dans les Petits Mondes Lexicaux, *Information Interaction Intelligence* vol. 4 - n°2.

HOLDEN, Clare J. & Russell D. GRAY. 2006. Rapid Radiation, Borrowing and Dialect Continua in the Bantu Languages. In '*Phylogenetic Methods and the Prehistory of Languages*', Cambridge: McDonald Institute for Archaeological Research, p.26.

KOELLE, S. W., 1854 [reprint 1963]. *Polyglotta Africana* ; Photomechanic Reprint of the Original Edition, Church Missiionary Society, London 1854. Freetown, Fourah Bay College, The University College of Sierra Leone.

KOROMA, R., 1994 : *Die morphosyntax des Gola*. Köln : Institut für Afrikanistik, Universität zu Köln.

NEW, B., PALLIER C., FERRAND L. & R. MATOS. 2001. Une base de données lexicales du français contemporain sur internet: LEXIQUE, *L'Année Psychologique*, 101, 447-462. http://www.lexique.org.

POZDNIAKOV, Konstantin. 1991. On the Mande and West-Atlantic groups: an approach to the quantitative comparative linguistics. *Mandenkan: bulletin semestriel d'études linguistiques mandé* 22, pp. 39-69.

POZDNIAKOV, Konstantin. 1993. *Sravnitel'naja grammatika atlanticeskix jazykov [Grammaire comparative des langues atlantiques]*. Moskva : Nauka.

POZDNIAKOV, Konstantin & Marie-Paule FERRY. 2001. Dialectique du régulier / irrégulier dans la reconstruction des classes nominales. in Nicolaï, R. (éd.), *Leçons d'Afrique, Filiations, ruptures et reconstitution de langues, Un Hommage à Gabriel Manessy*. Paris: Peeters.

POZDNIAKOV, Konstantin & Guillaume SEGERER. 2007. Similar place avoidance: A statistical universal. *Linguistic Typology* 11, pp. 307-350.

SAPIR, J. David. 1970. *Dictionnaire Jóola Kujamutay (Diola Fogny)*. Bignona. Disponible à l'adresse suivante : http://etext.lib.virginia.edu/african/Kujamaat/DIC/Joola-Dic.html.

SAPIR, J. David. 1971. West Atlantic : An inventory of the languages, their noun class systems and consonant alternations. *Current Trends in Linguistics 7* : 45-112. The Hague : Mouton.

SCHLENKER, C. F. 1864. *Grammar of the Temne Language*. London : Printed for the Church Missionary Society.

SCOTT, J. P. L., 1965. *An Introduction to Temne Grammar*. Sierra Leone : Government Printing Department.

SEGERER, Guillaume, 2002a. *La langue bijogo de Bubaque (Guinée Bissau)*. Paris, Louvain : Peeters (coll. Afrique et Langage, 3).

WILSON, William André Auquier. 2007. *Guinea Languages of the Atlantic Group : Description and Internal Classification*, Edited by Anne Storch. Frankfurt am Main : Peter Lang, Schriften zur Afrikanistik - Research in African Studies Vol. 12.

WILSON, William André Auquier. 1989. 'Atlantic', *in* J. T. Bendor-Samuel, *The Niger-Congo languages: a classification and description of Africa's largest language family*. Lanham MD, New York & London: University Press of America, pp. 81-104.

WINTZ Ed. 1909. *Dictionnaire français-dyola et dyola-français précédé d'un essai de grammaire*, Paris.

WESTERMANN, Diedrich. 1921. *Die Gola-Sprache in Liberia: Grammatik, Texte, und Worterbuch*. Hamburg : L. Friederichsen & Co.

WESTERMANN, Diedrich. 1928. *Die westatlantische Gruppe der Sudansprachen* (West-sudanische Studien, 5). Mitteilungen des Seminars für orientalische Sprachen, 31 (III. Abt.), pp. 63-86.