Brun-Trigaud Guylaine, Darlu Pierre, Gaillard-Corvaglia Antonella, Léonard Jean Léo, Sauzet Patric

EXPLORATION CLADISTIQUE DE L'ALLOC

INTRODUCTION
La cladistique comme systématique

L'objectif de cette contribution est résolument expérimental. Il s'agit d'explorer les structures diasystémiques du domaine languedocien à l'aide de la cladistique, qui est une technique de classification couramment utilisée par les généticiens et les biologistes. Le choix de cette méthode ne préjuge pas de l'isomorphie entre faits de langues et phénomènes de l'ordre du vivant, comme le voulait la linguistique de la deuxième moitié du 19ème siècle, quand August Schleicher reprenait les principes glanés au fil de la lecture de *l'Origine des espèces*, de Charles Darwin, pour fonder une théorie évolutionniste en linguistique génétique et typologique. L'objectif poursuivi ici est tout autre : il relève de la systématique, dans sa dimension diasystémique (Weinreich 1954), et il met à l'épreuve de la praxis cladistique les données dialectales d'un important atlas linguistique occitan, l'ALLOc (Ravier 1978), en exploitant les ressources d'une base de données fondée sur la compilation des données atlantographiques : le Thesoc.

Les données dialectales ont été choisies parmi les 1640 cartes et listes publiées de l'ALLOc, dont les données (187.500 entrées) ont été saisies au sein du *Thesaurus Occitan* (Thesoc : http://thesaurus.unice.fr/index.html). Afin de sélectionner les cartes destinées à construire une analyse cladistique, nous avons procédé en trois étapes. Dans un premier temps, le choix s'est porté, pour des raisons évidentes de comparabilité, sur les données dont les réponses provenaient toutes d'un même étymon lexical. Dans un deuxième temps, un tri a été effectué de manière à ce que les faits que l'on sait caractéristiques des parlers de ce domaine soient présents dans l'échantillon. Dans un troisième temps, nous avons cherché à établir une liste finale qui évite de donner indûment la primauté à certains phénomènes dont l'accumulation aurait faussé l'analyse en occultant d'autres phénomènes.

Les données extraites du Thesoc pour la présente étude constituent un corpus de 11 790 formes disponibles (90 cartes pour 131 localités), qui ont été intégrées dans une feuille Excel, après quoi les variables ont été extraites automatiquement. Chaque phénomène phonétique a fait l'objet d'une pondération dont les résultats figurent dans le tableau en annexe. L'aire étudiée est celle que couvre l'Atlas du Languedoc occidental (ALLOc) et elle correspond en effet assez justement à la partie occidentale de l'aire dialectale languedocienne. Le languedocien forme une aire centrale dans le domaine occitan, non pas en ce qu'il serait situé exactement au milieu géographique du domaine (il est plutôt méridional) mais en que cette aire se relie à toutes les autres aires dialectales de la langue : gasconne, limousine, auvergnate, vivaro-alpine et provençale. Sa centralité tient donc davantage de sa position de carrefour que d'un statut de méridien.

Les évolutions phonologiques dont nous avons eu soin de nous assurer qu'elles étaient représentées dans les formes soumises au classement sont de trois types. Il y a d'une part des phénomènes qui structurent largement l'espace occitan et qui individuent le domaine languedocien en soi. Il y a ensuite des traits qui opposent l'aire languedocienne à ses voisines. Ils sont présents dans les données étudiées ici parce que,

outre le caractère non absolu des limites dialectales et par un choix délibéré des auteurs de l'Atlas, l'aire couverte déborde le domaine dialectal languedocien vers le gascon et vers le nord-occitan, les voisins de cette partie occidentale du languedocien. Il y a enfin des phénomènes plus locaux, propres à diverses petites aires couvertes par l'ALLOc.

La méthode cladistique a été développée, à partir des principes élaborés par Hennig (1950,1966), pour rendre compte de l'évolution phylogénétique des espèces et reconstruire leur parenté. Le principe fondamental consiste à tirer parti de considérations sur l'évolution des caractères, en distinguant, pour chacun d'eux, une forme ancestral et une ou plusieurs formes dérivées. Deux espèces sont dites descendantes d'un même ancêtre commun lorsqu'elles partagent les mêmes états dérivées de ces caractères. En revanche, le partage de caractères dans leur forme ancestrale ne prouve pas nécessairement la parenté. Cette méthode est en parfaite adéquation avec la recherche des affinités linguistiques qui fonde l'apparentement entre deux langues ou entre deux formes dialectales, et leur rattachement à la même proto-langue, sur le fait qu'elles ont hérité de mêmes traits linguistiques dérivés.

Une fois ce principe posé, il s'agit de proposer une procédure qui permette de l'appliquer, le but étant d'obtenir une représentation arborescente des espèces (des langues ou des dialectes), qui rende compte de leurs liens de filiation. Dans un monde idéal, la construction d'un tel arbre, le « *perfect tree* » ou l'arbre parfait ne devrait poser aucune difficulté. Malheureusement les difficultés surgissent du fait que certains traits dérivés ont pu retourner à leur forme ancestral au cours de l'évolution (*réversion* ; les philologues utilisent le terme de *régression*), d'autres traits dérivés ont pu être acquis par convergence de manière indépendante chez deux espèces, et non par héritage, ou provenir d'un transfert d'une espèce à l'autre (transfert horizontal de gènes, emprunt dans le cas de langues). De tels événements, dénommés homoplasies, viennent perturber la véritable information sur les parentés entre espèces.

La procédure mise en place consiste donc à rechercher un arbre qui contiennent le moins d'événements homoplasiques. Cela revient, selon le principe de parcimonie, à minimiser le nombre de changements d'états de caractères dans l'arbre. Cette procédure suppose, en revanche, que les traits étudiées évoluent indépendamment les uns des autres et que les événements de réversions et de convergence soient rares[1].

Plusieurs programmes informatiques (Felsenstein, 1989, Swofford, 2002) permettent de mettre en application ces principes et de trouver l'arbre le plus parcimonieux, tout en tenant compte des diverses hypothèses que le biologiste (ou le linguiste) a pu formuler sur les modes d'évolution et de changement des caractères.

1. Protocole de recherche

L'application de la méthode cladistique aux données de l'ALLoc a nécessité les étapes suivantes :

1) La première étape consiste à construire les arbres diachroniques ou arbres de caractères qui configurent les changements phonologiques (voyelles et consonnes) des 90 mots extraits de l'ALLoc (Tableau 1). C'est l'étape où le linguiste définit ses hypothèses sur l'‘évolution des traits linguistiques. Il précise à la fois les connexions entre les différentes formes, le sens de leur transformation à partir de la forme dite ancestrale, et le poids attribué à chacune des transformations.

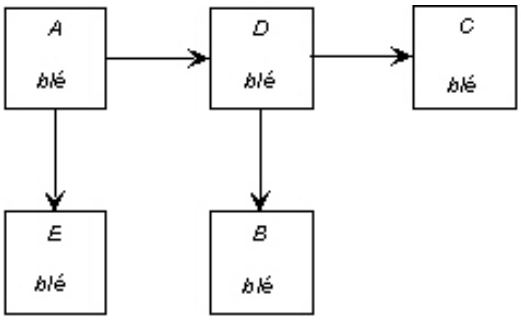
[1] Pour plus de détails, voir Darlu et Tassy (1993) ou Felsenstein (2004).

[1]	abeille	8	[31]	coeurduporc	3	[61]	nid	6
[2]	agneau	6	[32]	côte	3	[62]	noyé	9
[3]	agnelé	10	[33]	cou	7	[63]	œil	7
[4]	aiguille	6	[34]	cuir	7	[64]	œuf	4
[5]	aile	4	[35]	cuisse	11	[65]	oiseau	10
[6]	âne	3	[36]	cul	7	[66]	pain	6
[7]	arbre	5	[37]	doigt	3	[67]	peau	4
[8]	bâtir	3	[38]	dos	8	[68]	pie	8
[9]	blé	4	[39]	eau	5	[69]	pierre	1
[10]	bœuf	4	[40]	échelle	7	[70]	plier	6
[11]	bois	8	[41]	escalier	16	[71]	poirier	5
[12]	bouc	2	[42]	étoile	10	[72]	pondre	7
[13]	buis	6	[43]	farine	3	[73]	porte	1
[14]	chambre	7	[44]	feu	10	[74]	pré	1
[15]	chanvre	7	[45]	feuille	8	[75]	prés	4
[16]	chapeau	7	[46]	fiel	7	[76]	puits	11
[17]	châtaigne	8	[47]	fil	6	[77]	roues	8
[18]	chatte	5	[48]	foiedeporc	8	[78]	sac	8
[19]	chaud	2	[49]	froid	4	[79]	sauce	7
[20]	chauffer	5	[50]	fromage	10	[80]	sel	4
[21]	chemin	2	[51]	jambe	3	[81]	soleil	9
[22]	chemise	6	[52]	laine	5	[82]	sueur	4
[23]	cheval	10	[53]	lait	12	[83]	taupe	5
[24]	cheveu	7	[54]	loup	4	[84]	vache	5
[25]	chèvre	4	[55]	lune	6	[85]	veau	9
[26]	chier	4	[56]	marteau	3	[86]	vent	5
[27]	ciel	5	[57]	miel	3	[87]	verre	7
[28]	cire	5	[58]	mule	3	[88]	vert	3
[29]	ciseaux	3	[59]	nettoyer	8	[89]	vitre	5
[30]	clef	1	[60]	nez	1	[90]	voler	3

Tableau 1 : Les 90 mots-témoins de l'ALLoc et leur nombre de variantes phonologiques utilisés dans l'analyse cladistique

Dans le cas présent, les transformations ont été considérées comme irréversibles. Toutes les informations concernant la reconstruction de ces arbres diachroniques sont dans la section 3. Dans la Figure 1.I. est donné l'exemple des variations dialectales de « Blé », disposé en arbre de caractères – dont la forme est davantage un graphe qu'une arborescence à proprement parler.

2) La deuxième étape consiste en une codification de ces arbres diachroniques qui permette un traitement informatique sans perdre d'information. Il s'agit d'une procédure dite de « factorisation » illustrée par la figure F 1.II (programme *factor*, Felsenstein, 1989). Dans le cas de l'analyse des réflexes du mot-témoin « blé », les étapes des changements d'états entre A, B, C, D et E sont recodées en 0/1. Ainsi, dans la figure 1.II.. l'état B est recodé « 0110 », le deuxième index 1 de la série correspond au changement nécessaire pour passer de l'état A à l'état D (transformation A>D) et le troisième 1 pour passer de l'état D à l'état B (transformation (D>B), suivant le schéma F 1.I.



I

W	A->E x	A->D y	D->B z	D->C t
A	0	0	0	0
E	1	0	0	0
D	0	1	0	0
B	0	1	1	0
C	0	1	0	1

II

```
09.33QUER ... A ...
11.21PUIV ... E ...
47.10VLRE ... D ...
19.01SION ... B ...
24.11SVGM ... C ...
```

III

```
09.33QUER ... 0000 ...
11.21PUIV ... 1000 ...
47.10VLRE ... 0100 ...
19.01SION ... 0110 ...
24.11SVGM ... 0101 ...
```

IV

Figure 1 (F1). Etapes de l'analyse cladistique appliquée à l'ALLoc : Encodage de la matrice à partir du stemma diachronique

3) La troisième étape consiste à d'attribuer à chaque localité les formes phonologiques codées qui leur sont propres, pour les 90 mots. C'est ce que représente la figure F1.III, qui présente l'état de la forme « blé » dans quelques localités et la figure F1.IV où le descriptif de l'état est remplacé par sa forme telle qu'elle est codée en F1.II

4) Ces étapes permettent de construire une matrice avec en ligne chaque localité (n=40) et avec autant de colonnes qu'il y a de formes phonologiques différentes sur l'ensemble des 90 mots, soit 520 colonnes.

5) Le logiciel de parcimonie (PAUP*4.0b10, Swofford, 2001) recherche l'arbre (ou les arbres) minimisant le nombre de changements d'état de caractères, tout en tenant compte à la fois du poids attribué initialement à chacun des changements et de la contrainte d'irréversibilité. Cet arbre regroupe les localités qui

partagent les mêmes caractères phonologiques dérivés en différents « clades »[2] emboités.

6) Pour obtenir une indication de la robustesse de ces regroupements en clades, il est habituel de procéder à des procédures statistiques de ré-échantillonnage des données. Dans le cas présent, le ré-échantillonnage a été effectué par bootstrap (littéralement « chausse-pied », qui se réfère au procédé de ré-échantillonnage). Il consiste à créer *r* nouvelles matrices, chacune d'elles constituée des colonnes de la matrice initiale tirées au hasard avec remise. La recherche de l'arbre parcimonieux est ensuite exécutée sur ces *r* pseudo-matrices, conduisant à *r* nouveaux arbres. Sur cet ensemble d'arbres, sont calculées en pourcentage, le nombre de fois où sont retrouvés les mêmes clades. La valeur de ce pourcentage (proportion de bootstrap) indique le degré de robustesse du clade. Une forte valeur, par exemple une proportion de bootstrap de 90% signifie que ce clade est retrouvé sur 90% des arbres obtenus au cours du processus de ré-échantillonnage.

Maintenant que nous avons décrit la méthode et les principes qui fondent la procédure analytique, nous allons décrire les critères retenus pour établir les arbres de caractères qui alimentent la matrice. Nous présenterons ensuite les résultats dans la section 4.

2. Application au corpus de l'ALLoc

Cette contribution associe des coauteurs qui n'étaient pas précédemment investis dans la problématique cladistique mais qui ont voulu apporter leur connaissance des données occitanes et confronter leurs approches à cette démarche. La structuration de la diversité interne de l'espace occitan n'est pas en effet un domaine vierge et l'apport de la cladistique expérimenté ici dans sa mouture standard (c'est-à-dire à l'aide de matrices fondées sur des arbres de caractères) pourra d'autant mieux être mesuré qu'on saura comment le résultat est nourri de démarches et d'apports antérieurs de la phonologie historique de l'occitan et de la géographie linguistique. Les phénomènes généraux de différentiation phonologique du domaine occitan dont attestent les données sont :

- le traitement du groupe latin -kt- (et du groupe dérivé -gd-) qui distingue une large bande, en termes d'aire, correspondant au traitement par affrication (initialement) palatale, du Limousin à la Provence, du cluster occlusif hétéroganique, en incluant le centre et l'est du languedocien, tandis que le sud-ouest et une aire auvergnate conservent le premier stade évolutif par semi-vocalisation de la vélaire ([-jt-]) (ex *lach* vs. *lait*, *lèit* ; *freg* vs *freid* ([frˈet]),

- le bétacisme (*vin* [vˈi(n)] vs [bˈi(n)]), ample phénomène dont le foyer primitif est la Gascogne, où les documents les plus anciens l'attestent, mais qui a atteint au moins dès le début de l'époque moderne, un large tiers sud-ouest du domaine occitan,

- le traitement des obstruantes finales, au premier chef les occlusives, que l'aire qualifiée par Pierre Bec « d'arverno-méditerranéenne » tend à effacer alors que le sud-ouest « aquitano-pyrénéen » les conserve (Bec 1963) (*lop* [lˈu] vs [lˈup], *boc* [bˈu] vs [bˈuk]), l'aire étudiée présente aussi une bande intermédiaire

[2] Un *clade* est un *cluster*, groupe, un constituant de l'arborescence que forme un cladogramme, à savoir le résultat d'une analyse cladistique, ou systémique d'états dérivés à partir d'une forme prototypyique, appelée aussi « ancêtre » – l'étymon.

qui conserve les occlusives mais neutralise le point d'articulation au profit de la coronale ([lut], [but]),

- de manière partiellement corrélée au phénomène précédent et aussi au traitement d'-s en coda (cf. nord-occitan vs languedocien), la marque de pluriel, -s, s'efface ou se trouve remplacée par une marque vocalique ou quantitative au nord et à l'est (*ròdas* [rːɔdaːˑ] vs [rːˈɔðɔs]),
- la labialisation de -a- prétonique dessine une vaste aire dans le Massif Central qui s'étend à la fois en domaine languedocien et nord occitan (*farina* [fɔɾˈinɔ] vs [fɔrˈinɔ]).

Puisque le domaine envisagé est le domaine languedocien occidental, les aires dialectales adjacentes dont on peut voir apparaître les spécificités aux marges de la zone étudiée sont le nord occitan d'une part et le gascon d'autre part.

Du nord-occitan on peut trouver dans l' ALLOc les traits suivants :

- la palatalisation (suivie d'autres évolutions) des vélaires devant -a (*chabra* [ʧˈabɾɔ], [tsˈabɾɔ], [sˈabrɔ] vs *cabra*[3]), phénomène caractéristique de l'aire nord occitane, qui ne se manifeste que très marginalement dans nos données,
- la réalisation palatale de la fricative coronale (initialement alvéolaire) : *sac* [ʃˈak] vs [sˈak],
- l'effacement de -d- intervocalique latin (*suar* vs *susar*),
- la perte (avec allongement compensatoire éventuel) de l'-s en coda (*tèsta* [tˈeːtɔ], [tˈejtɔ] vs [tˈɛstɔ]), évolution qui retentit sur le marquage du pluriel comme on l'a déjà vu,
- l'aphérèse de l'a- initial (*'belha* en face d'*abelha*).

Le gascon s'annonce sur la bordure ouest de l'espace étudié par les propriétés suivantes :

- la débuccalisation de la fricative labio-alvéolaire sourde -f, trait prototypique s'il en est du gascon (*haria* vs *farina*),
- la chute de -n- intervocalique (que le même exemple *haria* vs *farina* illustre),
- la conservation de -*n* final, que l'on retrouverait à l'est en Provence, mais qui localement oppose le gascon autant au languedocien occidental qu'au limousin (*vin* [bˈĩŋ] vs [bˈi], [vˈi], *pan* [pˈãŋ] vs [pˈa]),
- le traitement de la latérale géminée -ll- en [-ɾ-] et en [-t] (*vedèl* vs *vedèth*), qui est un autre trait caractérisant du domaine gascon,
- la réalisation du produit du suffixe -ariu- (*perèr* vs *perièr*).
- les faits de métathèse des rhotiques, très lexicalisés mais dont le foyer d'intensité maximale est la Gascogne (*cramba, craba* vs *cambra, cabra*).

[3] L'illustration des phénomènes présents dans les données est faite en utilisant une notation orthographique d'usage quand elle est suffisamment explicite (formes en italiques : *cabra, chabra*) et la notation phonétique au besoin ([kˈabrɔ], [ʧˈabrɔ], [tsˈabrɔ]..).

- la réalisation centralisée de -a post-tonique (*luna* [lˌynə] vs [lˌynɔ] réalisation majoritaire), qui est globalement un trait d'ouest gascon mais vient dans l'Entre-deux-mers au contact du languedocien,

Quand aux phénomènes plus locaux on relèvera sans pouvoir être exhaustif :

- la surévolution et le blocage ou la régression de diphtongaisons conditionnées occitanes (*uèlh* ['ɛl] vs [ʮ'ɛl] et *fòc* vs *fuòc* [fj'ɔk],
- la réalisation moyenne de -u- (*luna* [lœnɔ] pour [l'ynɔ]),
- la dépalatalisation des affriquées dans une partie du languedocien (*fach* [f'ats] pour [f'aʧ]),
- la diphtongaison de *i* suivi de latérale (*fiel, fial* pour *fil*),
- la diphtongaison (récente au sens où elle est distincte de la diphtongaison dite *romane* de cette voyelle) de *ò* tonique qui apparaît en languedocien montagnard (comme par ailleurs sous une forme plus évoluée en Provence martime) *pòrc* [pw'ɔrk] pour [p'ɔrk],
- la palatalisation et la réalisation interdentale et éventuellement obstruante de la latérale issue d'une géminée latine issue d'une géminée latine propre à

certaines aires pyrénéennes (*lhuna* [ʎ'unɔ], [ļ'ynɔ], [θ'ynɔ])

Dans la construction des stemmas qui ont servi de base à l'établissement des regroupements en arbres de caractères et au traitement cladistique, nous avons organisé les formes en tenant compte du schéma évolutif supposé en partant du principe, discutable certes mais qui a le mérite d'être une hypothèse de travail assumée aussi bien par nous que par une longue tradition en géolinguistique, que les formes actuelles observables dans l'espace reflètent des étapes de l'évolution dans le temps, ce qui a demandé sans cesse de faire des choix. Il a fallu trancher entre plusieurs solutions possibles afin d'aboutir à un objet qui soit le plus manipulable possible. Les stemmas construits (ou *arbres de caractères*) sont en effet des graphes strictement arborescents alors que l'évolution phonologique ne l'est pas. Face à quatre formes [l'ynɔ], [l'yɔ], [l'ynə], [l'yə][4] il faut en effet décider si l'on opte d'abord pour un embranchement « n→ Ø » vs « n conservé », ou d'abord « a → ə » *versus* « a → ɔ ». On ne peut représenter un processus évolutif organisé en ondes successives, faisant intervenir la centralisation du continuateur d'-a final atone en recouvrement et en débordement partiels du domaine touché par l'onde de la chute d'-n intervocalique. En un sens, le choix de placer une évolution avant l'autre n'est pas d'une importance capitale dans un premier temps : ce n'est qu'une affaire d'exposition. En revanche, il importe de conserver une pondération constante des changements. C'est en effet la pondération des « pas » dans les branches du stemma qui importe en termes de conception de l'évolution diachronique. Cette pondération a été effectuée, en ce qui concerne notre corpus, sur la base de deux critères : la distance phonétique brute d'une part, le caractère ou non phonologique du changement d'autre part.

[4] Ces quatre formes, choisies pour la simplicité de l'exposition, ne sont pas toutes dans nos données (il y manque [l'yɔ] attesté un peu plus loin) mais il y a des exemples équivalents et on a tenu compte de cette problématique dans le classement des formes issues de LUNA en ne faisant pas de la différence pourtant récurrente dans le traitement de la voyelle finale un embranchement basique du graphe.

Ainsi dans le graphe évolutif ou stemma (ou *arbre de caractères*) de LUNA, *luna*, la palatalisation de la latérale initiale, *lhuna* qui va avec celle de la latérale géminée, *galhina*, reçoit un fort coefficient en tant qu'elle peut introduire une neutralisation phonologique (ou plutôt qu'elle conserve l'opposition phonologique de la latérale forte, initiale ou géminée et de la latérale faible). Les surévolutions spectaculaires de *lhuna* ([l̥ˈynɔ], [θˈynɔ) n'ont reçu qu'un coefficient faible en tant que changements phonétiques seulement, mais ce poids faible se cumule avec le contraste de fondamental, qu'il renforce dans sa différenciation avec les parlers du type commun (*luna*, *galina*), avec rétention de la latérale étymologique.

La variation de la voyelle finale atone n'a reçu qu'un faible coefficient. La réalisation [ɔ] a été placée au centre du schéma, bien que la série réelle des étapes phonétique soit sujette à débat. Le placement en périphérie du diagramme des formes à voyelle centrale se justifie en occitan par le fait que la réalisation labialisée placée au centre s'accompagne elle du maintien de l'opposition *-e* [e], *-a* [ɔ] alors que la centralisation peut aller de pair avec une réduction de l'inventaire phonologique des voyelles post-toniques *-e* [ə] comme *-a* [ə].

Enfin une ultime étape a consisté à choisir 40 localités parmi les 131 initiales : les points d'enquêtes ont fait l'objet d'un regroupement par 3 ou 4 en fonction des affinités phonétiques qu'ils présentaient entre eux, puis un point a été choisi comme le plus représentatif du groupe, en écartant ceux qui étaient trop "singuliers" sans toutefois opter pour le trop "consensuel".

3. Résultats : une tripartition et huit aires

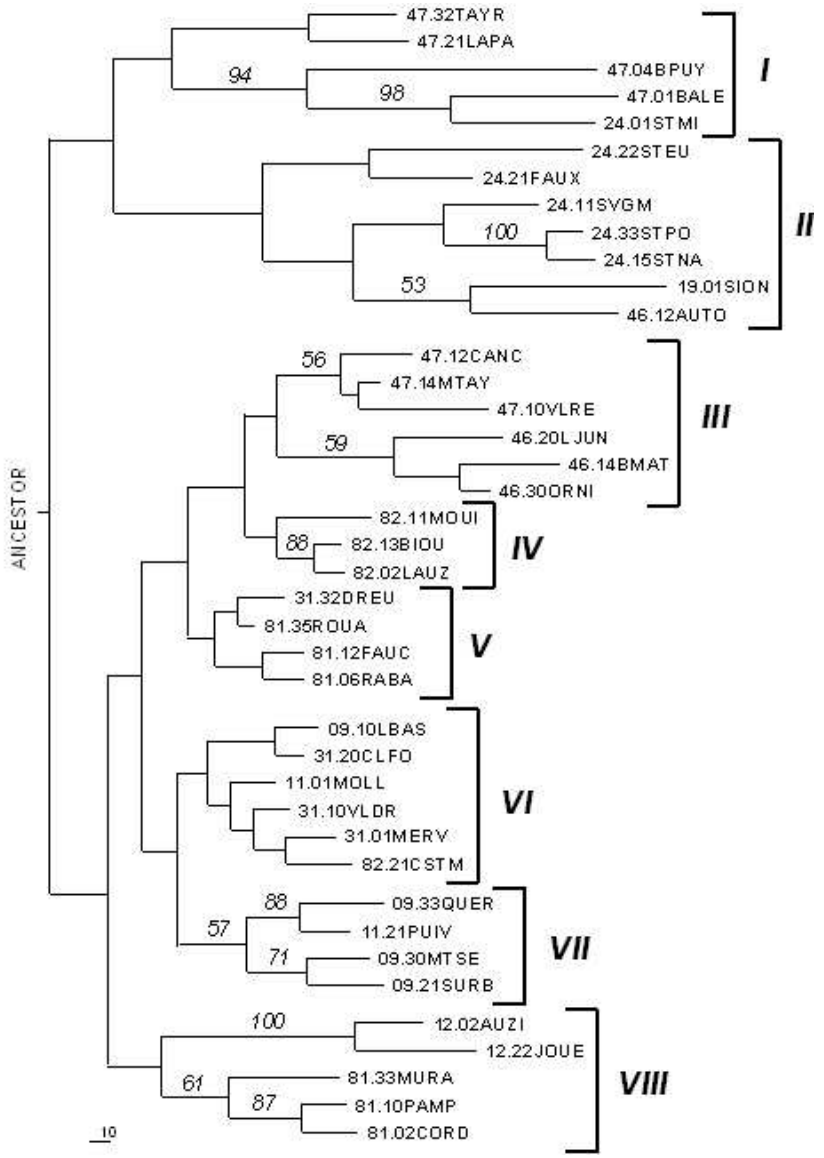
L'arbre parcimonieux obtenu par PAUP*4.0 est unique (2833 pas) avec un indice de cohérence de CI= 0.37 et un indice de rétention de RI=0.72. Ces indices rendent compte du degré d'homoplasie (en l'occurrence, ici, les seules convergences, c'est-à-dire les changements d'état de caractères sur deux ou plusieurs branches distinctes). Formellement, CI= R/L où R est le minimum de changements auxquels on s'attend, compte tenu des données, dans le cas d'un « arbre parfait », et L le nombre de changements effectivement trouvés sur l'arbre parcimonieux. RI, indice de rétention, est une normalisation du CI pour se distribuer dans l'intervalle entre 0 et 100%. Plus les valeurs de CI ou RI sont élevés, plus on se rapproche de « l'arbre parfait » et donc du constat d'absence d'homoplasie.

La figure F2 représente l'arbre de parcimonie. Les longueurs de branches sont proportionnelles au nombre de changements qui les définissent (l'unité d'échelle est de 10 changements). Les valeurs portées sur les branches de l'arbre correspondent aux proportions de ré-échantillonnage par bootstrap supérieures à 50% (*r* = 200 répétitions). Un ré-échantillonnage par Jackknife 50% donne des proportions équivalentes[5].

[5] Le Jackknife ou « couteau suisse » est une autre façon de procéder à un ré-échantillonnage. Ceci dit, cette procédure alternative ou complémentaire de ré-échantillonnage est considérée comme plus logique quand les données ne sont pas soutenues par un modèle probabiliste, comme c'est le cas ici, dans le cadre d'une analyse phonologique. Cette procédure a cependant le désavantage d'être dépendante du choix que l'on fait de la proportion de caractères qui seront tirés au hasard à chaque ré-échantillonnage (50%, 20 %,

Il faut signaler les faibles valeurs de proportion de boostrap et les branches globalement plus longues pour les communes de Dordogne et du Lot-et-Garonne. Les traits épais délimitent les deux grands clades.

La figure F1 présentée supra dans la section 2 décrivait comment l'on passe des stemmas, ou graphes de dérivation d'états diachroniques (en F1.I) à la matrice de caractères indexant les « pas » qui rassemblent les localités (F1.II), auxquelles sont attribués des états structuraux ou caractères (F.1.III), et des séquences de pas indexés dans les chaînes évolutives attestées (F.1.IV). La carte de la figure 3 (F3, infra) localise les 40 communes, en les regroupant selon les 8 clades de l'arbre de la figure F2 (numérotés I à VIII). Les traits épais délimitent les grands regroupements de clades « frères ». Les traits reliant deux communes indiquent leur regroupement en un clade dont la %BP est supérieure à 70%.



90%, etc.). Toutes ces mesures de ré-échantillonnage ont pour objectif de tester statistiquement la robustesse du résultat de l'analyse cladistique.

Figure F2. Résultats de l'analyse cladistique : configuration des huit clades.

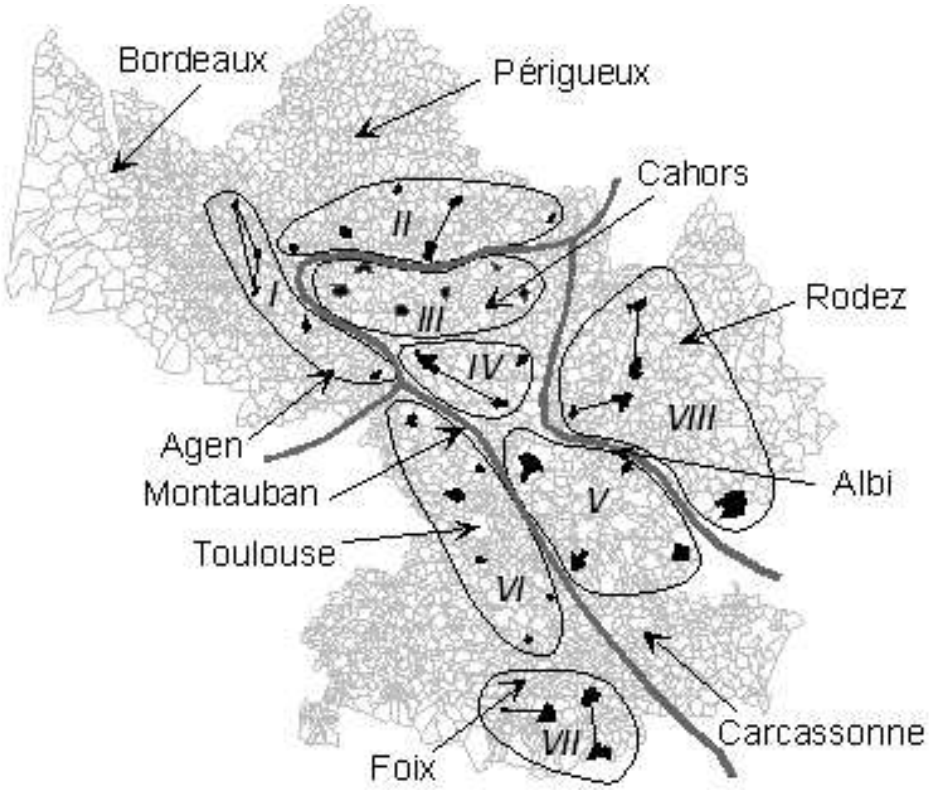


Figure F3. Résultats de l'analyse cladistique, en huit clades

L'objet de connaissance produit par cette analyse donne une image aisément interprétable de la structuration du diasystème languedocien dans son contexte occitan. On voit une aire centrale fragmentée en trois sous-aires – un languedocien central qui se subdivise en *quercinois* (Lot et Tarn-et-Garonne, en *albigeois* (Tarn) et en *rouergat* (Aveyron) –, dont l'enveloppe, resserrée à la hauteur de Montauban, forme un couloir en carafe qui va de Carcassonne à Cahors. Au nord-ouest de cette aire centrale tripartite apparaît ce qu'on peut caractériser comme *occitan pré-septentrional girondin*. Ailleurs, à l'ouest et à l'est, l'aire centrale est flanquée d'une bande constituée de trois aires, distribuées entre Foix au sud, Toulouse au centre et Agen au nord-ouest d'une part, et de l'aire de Rodez au centre-est.

CONCLUSION ET PERSPECTIVES

La cladistique n'est qu'une méthode parmi d'autre de classification des langues, que l'on peut même considérer, en dialectologie, comme complémentaire d'autres méthodes de taxinomie des aires et des systèmes dialectaux comme la dialectométrie (Goebl 1982[6]).

Elle présente ceci de particulier qu'elle fonde la classification sur des séries dérivées, ou des chaînes de dérivations, pondérées en fonction de critères structuraux –

[6] Voir également le site Internet http://www.dialectometry.com/

autrement dit, une *valeur*, un poids est attribué aux changements phonétiques. A ce titre, la cladistique s'avère être un outil particulièrement adapté à la dialectologie structurale, voire un outil d'exploration de ce que Weinreich, dans son article fondateur (*op. cit.)* appelait le *diasystème*, en tant que métasystème qui subsume la variation, et permet de restituer une totalité structurale par la mise en équivalence et en corrélation de ses diverses parties, infirmant l'hypothèse de Paul Meyer et de Gaston Paris sur la non existence des dialectes et sur la primauté du continuum dialectal. Toutes les méthodes taxinomiques, aussi bien la dialectométrie que la cladistique, ont pour objectif de définir des entités dialectales pertinentes sur le plan systémique et structural, sans préjuger des catégorisations externes (historiques, épilinguistiques), et sans pour autant remettre en cause l'unité d'un domaine linguistique, puisque les parties discrètes sont les composantes organiques d'une totalité organisée et hiérarchisée, qui répartit ses variables et ses contraintes dans un champ structural. Les résultats de la carte de la figure 3 peuvent certes présenter un grand intérêt pour les géographes et les historiens, ou les amateurs de « terroirs ». Il n'en reste pas moins que la méthode cladistique, par la multiplicité de ses outils de mesure de la cohérence systémique, manifeste avant tout une attitude de recherche falsifiable et prudente : elle permet de définir des configurations et des structures internes entre sous-systèmes ou dialectes et sous-dialectes d'un réseau de points d'atlas linguistique, de manière à la fois parcimonieuse dans le séquençage des états dérivés ou supposés tels, et plausible en termes de probabilités et de proportions. Nous laissons aux spécialistes d'autres domaines de recherches sur l'histoire et la géographie du monde occitan tirer les conclusions auxquelles il leur semblera loisible d'aboutir à la lecture des résultats présentés dans les figures 2 et 3. Ils ne devront cependant oublier à aucun moment la nature sémiotique des entités linguistiques – les langues ne sont pas des organismes vivants, mais des agencements de symboles contraints par le lexique et la grammaire, sur une polarité qui va de la cognition humaine aux interactions entre communautés historiques. La leçon de la cladistique, appliquée aux données de l'ALLOc, réside sans doute davantage dans les contraintes de la méthode (séquençage des caractères, ou états structuraux en relation d'implication mutuelle, construction des matrices, tests de probabilité et de re-échantillonnages), que dans les aires délimitées sur la carte. Il importe, une fois de plus, de rappeler que la carte n'est pas le territoire, mais une représentation raisonnée de celui-ci.

REFERENCES

Bec Pierre 1963, *La langue occitane*, Paris, P.U.F.

Darlu Pierre, Tassy Pascal 1993. *La reconstruction phylogénétique. Concepts et méthodes*, Paris, Masson, (http://sfs.snv.jussieu.fr/pdf/Darlu_Tassy_online.pdf.).

Felsenstein Joseph 1989, PHYLIP « Phylogeny Inference Package (Version 3.2) », *Cladistics* 5: 164-166.

Felsenstein Joseph 2004. *PHYLIP. Phylogeny Inference Package*, version 3.6b. Distributed by the author. Department of Genome Sciences, Seattle, University of Washington.

Goebl Hans 1982, *Dialektometrie. Prinzipien und Methoden des Einsatzes der numerischen Taxonomie im Bereich der Dialektgeographie*, Vienne, Verlag der Öst. Akademie der Wissenschaften.

Hennig Willi 1950, *Grundzüge einer theorie der Phylogenetischen systematik*, Berlin, Deutscher Zentralverlag.

Hennig Willi 1966, *Phylogenetic Systematics*. Urbana, University of Illinois Press.

Swofford David, 2002. *PAUP*. Phylogenetic Analysis Using PArsimony* (*and other methods), version 4.0, Sunderland, Massachusetts, Sinauer Associates.

Ravier Xavier 1978, *Atlas linguistique et ethnographique de la France Languedoc occidental*, Paris, CNRS.

Weinreich Uriel, 1954. « Is a structural dialectology possible? », *Word* 4: 388-400.

Tort Patrick 1980. *Evolutionnisme et linguistique*, suivi de *August Schleicher La théorie de Darwin et la science du langage ; de l'importance du langage pour l'histoire naturelle de l'homme*, Paris, Vrin.