La typologie aréale comme fenêtre sur la préhistoire de l'Afrique subsaharienne

Dmitry Idiatov

Institut National des Langues et Civilisations Orientales

École doctorale n°265

Langues, littératures et sociétés du monde

Habilitation à diriger des recherches

présentée par

Dmitry IDIATOV

soutenue le 29 juin 2023

La typologie aréale comme fenêtre sur la préhistoire de l'Afrique subsaharienne

Volume 3 : Mémoire de synthèse

Garant :

M. Guillaumes JACQUES	Directeur de recherche, CNRS

MEMBRES DU JURY :

M. Guillaumes JACQUES	Directeur de recherche, CNRS
Mme Sonia CRISTOFARO	Professeur des universités, Sorbonne Université
M. Jeff GOOD	Professeur, University at Buffalo
Mme Tatiana NIKITINA	Directrice de recherche, CNRS
M. Erich ROUND	Professeur, University of Surrey
M. Didier DEMOLIN	Professeur émérite, Université Sorbonne Nouvelle - Paris 3

Table des matières

Abréviations

1, 2, 3	première, deuxième et troisième personne
BEN	bénéfactif
COP	copule
DAT	datif
DEF	défini
DEM	demonstratif
EMPH	emphase
F_{LV}	fréquence estimée des occlusives labiales-vélaires dans une langue
FOC	focalisation
FUT	futur
GAM	modèle additif généralisé / modélisation additive généralisée
INDF	indéfini
IPFV	imperfectif
LOG	logophorique
LV	labiale-vélaire
MNFP	marque de négation en fin de phrase
NEG	négation
NONHUM	non humain
OBJ	objet
PFV	perfectif
PID	pondération inverse à la distance
PL	pluriel
POSS	possessif
PQ	question totale
PRF	parfait
PROG	progressif
PROP	nom propre
PST	passé
QUO	marque de citation
REFL	réfléchi
REL	relatif
SBJ	sujet
SBJV	subjonctif
SG	singulier
STAT	statif

TAM temps, aspect, mode

1 Introduction

Ce mémoire de synthèse résume mes recherches des dernières années en typologie aréale des langues de l'Afrique subsaharienne. Dans le cadre de mes recherches sur les langues africaines, je me suis servi de la typologie aréale comme d'un outil supplémentaire et puissant pour découvrir la préhistoire linguistique de l'Afrique subsaharienne. La préhistoire linguistique ici est comprise à la fois dans le sens de la préhistoire des systèmes linguistiques eux-mêmes, c'est-à-dire la reconstruction d'aspects des systèmes linguistiques concernés, et dans le sens de la préhistoire des populations de la région comme elle se révèle par le biais des langues parlées par ces populations, par exemple concernant l'histoire des mouvements migratoires des populations, des périodes de contacts entre les populations parlant des langues différentes ou des évènements de changement de langues parlées par les populations de la région.

En guise d'introduction à la préhistoire linguistique de l'Afrique subsaharienne, je commence la discussion en présentant dans le chapitre 2 le soit disant « paradoxe africain », le terme qui réfère à la disparité surprenante entre l'ancienneté de la présence de l'homme en Afrique, le berceau de notre espèce, et le faible niveau présumé de la diversité linguistique sur ce continent, puisque plus de 2000 langues africaines sont supposées appartenir à une de quatre macro-familles seulement, à savoir le niger-congo, le nilo-saharien, l'afro-asiatique et le khoisan. Bien que ce degré extrême d'homogénéité généalogique des langues africaines soit de plus en plus mis en doute, il reste toutefois également clair que, dans une large mesure, elle reflète quand-même la réalité linguistique du continent, surtout en ce qui concerne les domaines niger-congo et afro-asiatique. Il s'ensuit que plusieurs événements de grande expansion d'un nombre limité de groupes linguistiques ont dû avoir lieu sur le continent. En Afrique subsaharienne, il s'agit principalement de l'expansion des langues niger-congo, qui occupent actuellement la plus grande partie de cette partie du continent. En même temps, un examen plus approfondi du phylum niger-congo révèle qu'il est caractérisé par un degré important de diversité interne et que cette diversité a une organisation spatiale non triviale indépendante de la sous-classification généalogique de ce phylum. Cette organisation spatiale dans un nombre d'aires linguistiques peut être étudiée fructueusement dans le cadre de la typologie aréale. En captant les signaux qui peuvent être fournis par la distribution spatiale inégale de certains traits linguistiques, la typologie aréale peut au minumum contribuer à une meilleure compréhension de l'histoire de la diffusion de ce phylum et à une réévaluation critique des reconstructions

qui ont été proposées dans la littérature, et à maximum, potentiellement également révéler des traces des langues des populations pré-niger-congo.

Dans le chapitre 3, je présente les deux objectifs principaux de la typologie aréale, à savoir l'objectif descriptif et l'objectif explicatif, qui sont logiquement hiérarchisés (§3.1), ainsi que mon approche méthodologique (§3.2-3.3). L'intérêt principal de la typologie aréale pour moi est son potentiel explicatif et historique. Dans cette perspective, force est de constater que pour capter au mieux le signal diachronique dans les données synchroniques et arriver ainsi à une explication diachronique plus plausible, il faut tenir compte au maximum de toute la diversité dans les données disponibles. De ce simple constat découlent les principes de base qui déterminent mon approche méthodologique de la typologie aréale que je discute dans §3.2. Ainsi, c'est une approche bottom-up qui départ d'une hypothèse de recherche claire formée par une connaissance approfondie des langues de la région et fait appel aux big data faisant toutefois attention au principe analytique général *garbage in, garbage out* et essayant dans la mesure du possible d'éviter d'imposer des catégories arbitraires aux données qui sont par leur nature graduelles. Dans l'objectif explicatif, cette approche préconise des explications en termes de scénarios qui impliqueraient des mécanismes concrets de changement linguistique. Je montre comment on peut traduire ces principes méthodologiques de base en pratique dans §3.3 en commençant par la collecte, le nettoyage et la mise en forme des données (§3.3.1) et en procédant à leur analyse spatiale et visualisation (§3.3.2) avec des méthodes aussi rudimentaires et peu efficaces que les motifs de points (§3.3.2.1), des méthodes déterministes plus sophistiquées comme l'interpolation spatiale (§3.3.2.2) et des méthodes statistiques plus performantes et plus rigoureuses comme la modélisation additive généralisée (§3.3.2.3).

Les deux derniers chapitres 4 et 5 présentent deux études de cas en typologie aréale des langues de l'Afrique subsaharienne qui illustrent l'application de mon approche méthodologique et son potentiel explicatif à deux types de traits linguistiques très différents. La première étude de cas qui fait l'objet du chapitre 4 (basé sur Idiatov 2018) présente la typologie aréale d'un trait morphosyntaxique, à savoir la négation en fin de phrase, qui constitue une particularité typologiquement frappante du marquage de la négation dans les langues de différentes régions de l'Afrique subsaharienne. La deuxième étude de cas qui fait l'objet du chapitre 5 (basé sur Idiatov & Van de Velde 2021)[1] présente la typologie aréale d'un trait phonologique typologiquement encore plus exclusive pour l'Afrique subsaharienne, à savoir les occlusives labiales-vélaires, telles que *k͡p*, *g͡b* et *ŋ͡m*. L'aspect particulièrement novateur de cette étude en typologie aréale

[1] En tant que premier auteur de cette publication, j'ai élaboré sa méthodologie, j'ai constitué le corpus de données et j'ai fait l'essentiel de l'interprétation des résultats. Le deuxième auteur a contribué à la rédaction de l'article et a fait des commentaires détaillés à toutes les étapes de la recherche.

est qu'elle considère les occlusives labiales-vélaires dans la perspective de la profondeur de leur ancrage dans les langues où elles sont attestées en se servant des estimations de leurs fréquences lexicales en tant qu'approximation. Les deux études montrent bien qu'il est possible d'extraire de l'information sur la préhistoire linguistique de certains traits typologiques et de formuler des hypothèses détaillées concernant les routes migratoires préhistoriques des populations. Evidemment, les différents traits linguistiques peuvent être plus ou moins informatifs pour la reconstruction de la préhistoire linguistique et que la profondeur préhistorique à laquelle les différents traits linguistiques nous permettent d'accéder peut également varier d'un trait à l'autre. Des deux traits étudiés ici, il est clair que c'est la fréquence lexicale des occlusives labiales-vélaires qui est le plus informatif pour la reconstruction de la préhistoire plus ancienne de l'Afrique subsaharienne et qui nous permet de repérer des traces linguistiques des populations substrat pré-niger-congo. Les deux études de cas soulignent en même temps qu'il pourrait y avoir de relations causales (parfois, pas du tout triviales) entre divers traits linguistiques qui à première vue seraient sans aucun rapport et qu'un modèle explicatif plus complexe pourrait s'avérer plus adéquat. Ainsi, une relation préférentielle entre les occlusives labiales-vélaires et les parties expressives du vocabulaire, d'un côté, et la position initiale du radical, de l'autre, peut sembler à première vue aléatoire, mais ne l'est pas en réalité et peut largement être expliquée à travers le phénomène de la prosodie d'emphase-C, la proéminence prosodique des consonnes initiales du radical dont le corrélat phonétique typique est la longueur de la consonne.

2 Le paradoxe africain et les aires linguistiques

Comme l'a souligné très récemment Hammarström (2018:24–25), il y a étonnamment peu de traces de la diversité linguistique aui a dû exister en Afrique de l'Ouest avant l'expansion de la famille niger-congo. Ce n'est qu'un aspect d'un paradoxe qui concerne l'ensemble de l'Afrique : le continent d'origine d'Homo Sapiens est celui où l'on s'attend à trouver la plus grande diversité linguistique, et pourtant, jusqu'à récemment, suite à la classification influente de Greenberg (1963), les langues africaines étaient largement supposées appartenir à une de quatre macro-familles seulement, à savoir le niger-congo, le nilo-saharien, l'afro-asiatique et le khoisan (Hombert & Philippson 2009:146). L'attention portée au paradoxe africain s'est traduite par trois grands axes de recherche récents : (i) une quête d'identification des isolats linguistiques ; (ii) un renouveau d'intérêt pour la classification des langues, suite à la reconnaissance croissante du fait que l'unité généalogique des quatre phylums linguistiques africains de Greenberg est une hypothèse qui reste en grande partie à vérifier ; et (iii) l'attention portée aux grandes aires linguistiques.

Les isolats linguistiques sont des langues pour lesquelles on ne peut pas identifier de langues généalogiquement apparentées. La raison de l'intérêt pour les isolats linguistique en Afrique est que ces isolats peuvent être considérés comme des traces de familles de langues qui ont été anéanties par l'expansion des langues niger-congo en particulier. Jusqu'à présent, les seules autres traces linguistiques de ces familles disparues ont été recherchées dans les lexiques spécialisés des langues contemporaines, notamment dans les domaines de la chasse, de la cueillette et de la pêche. Grâce au lien arbitraire entre le sens et la forme des éléments lexicaux, on peut être relativement certain que les ensembles de mots désignant des concepts liés à l'écologie de la région où est parlée une langue contemporaine et qui n'ont pas d'équivalent dans les langues apparentées, ont été empruntés ou hérités d'une population d'origine. Ce n'est pas le cas pour les traits typologiques tels que ceux utilisés pour caractériser les aires linguistiques, car ils peuvent apparaître indépendamment et se répandre à travers les langues. Toutefois, je propose qu'il est possible d'extraire de l'information sur la préhistoire linguistique de certains traits typologiques et de les mettre éventuellement en relation avec des populations substrat pré-niger-congo si l'on ne se limite pas au simple constat que le trait est présent dans une aire linguistique présumée et absent dans les langues généalogiquement proches en dehors de cette aire, mais si l'on étudie plutôt la distribution spatiale de l'enracinement du trait typologique, dont les implications

historiques peuvent alors être interprétées en relation avec les caractéristiques climatiques et géographiques des zones où le trait en question est particulièrement prédominant. L'hypothèse du substrat serait alors le résultat de l'examen et de l'élimination de tous les scénarios alternatifs qui auraient pu conduire à la distribution actuellement observée. Par ailleurs, il a été démontré que les lexiques spécialisés empruntés ou hérités des langues pré-niger-congo peuvent être caractérisés par des traits typologiques de leur langue source. Ainsi, les clics dans les langues bantu sud telles que le gciriku [gcir1234], le mbukushu [mbuk1240] et le kwangali [kwan1273] sont particulièrement fréquents dans les termes de pêche (Fisch 1984).[2]

Un récent aperçu critique des progrès réalisés dans la classification des langues africaines peut être trouvé dans Güldemann (2018a). Le sous-groupement généalogique du phylum niger-congo est compliqué par le grand nombre de langues qui en font partie, par des lacunes dans notre connaissance d'un nombre important de ces langues et par certains aspects des systèmes phonologiques et morphologiques de ces langues. L'appartenance de certaines familles linguistique au phylum niger-congo reste également à confirmer. Bien que la typologie aréale en elle-même ne peut pas résoudre directement les problèmes de classification, elle peut contribuer à une meilleure compréhension de l'histoire de la diffusion de ce phylum qui occupe en lui seul la plus grande partie de l'Afrique subsaharienne, et à une réévaluation critique des reconstructions qui ont été proposée dans la littérature. Ainsi, la typologie aréale peut démontrer qu'il est possible qu'un trait typologique soit ancien dans la région où il se trouve tout en étant récent dans les familles linguistiques dans lesquelles il est actuellement réalisé. Un bon exemple d'un tel trait typologique est la présence des occlusives labiales-vélaires dans beaucoup de langues de l'Afrique subsaharienne septentrionale qui ont souvent était reconstruites pour diverses familles de langues niger-congo et pour le proto-niger-congo lui-même (cf. Cahill 2017; 2018) suite à l'observation correcte mais historiquement non pertinente que ces occlusives sont attestées dans de nombreuses langues actuelles de ces familles.

Les aires linguistiques sont des zones qui présentent un regroupement remarquable de traits linguistiques dont la présence ne peut pas être attribuée à un héritage commun, car ces traits traversent les frontières généalogiques. L'existence de grandes zones linguistiques en Afrique a été signalée pour la première fois par Greenberg (1959) et Larochette (1959), puis reprise par Meeussen (1975) et Heine (1975), et ensuite largement ignorée, jusqu'à ce qu'un volume édité par Heine & Nurse (2008) relance l'intérêt pour le sujet. Le chapitre de Güldemann (2008), mis à jour dans Güldemann (2010; 2018b), dans lequel il reprend l'idée de Greenberg d'une zone centrale africaine

[2] Le code composé d'une combinaison de quatre lettres et de quatre chiffres entre crochets après le nom d'une langue est son identifiant Glottolog (Hammarström et al. 2022).

et la renomme *Macro-Sudan Belt* (voir déjà Güldemann 2003), a été particulièrement influent. Il fournit une liste de traits aréaux phonologiques (dont les implosives, les occlusives labiales-vélaires et les inventaires vocaliques étendus [7+]) et morphosyntaxiques (dont la logophoricité et les marques de négation en fin de phrase), qui étaient déjà connues séparément dans la littérature. La zone Macro-Sudan Belt est représentée par le numéro III sur la figure 1 des aires linguistiques de l'Afrique, d'après Güldemann (2018b:473). L'étendue géographique de la Macro-Sudan Belt de Güldemann correspond en grande partie à une aire linguistique reconnue par Clements & Rialland (2008) sur la base de traits purement phonologiques, qu'ils appellent la *Sudanic zone*, caractérisée par la présence de battues labiales, d'implosives et d'occlusives labiales-vélaires, de voyelles nasales et d'une harmonie « ATR ». Ce dernier trait de l'harmonie « ATR » fait l'objet d'une étude récente par Rolle et al. (2020) qui démontrent que dans les langues de la Macro-Sudan Belt une relation antagoniste existe entre l'harmonie « ATR » et la présence des voyelles intérieures dans les inventaires vocaliques de sorte que les langues avec l'harmonie « ATR » se trouvent regroupées dans trois régions discontinues.

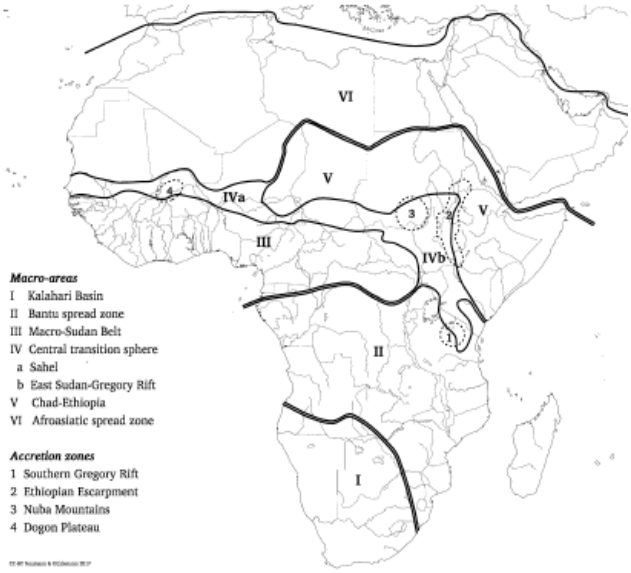


FIGURE 1. Les aires linguistiques en Afrique selon Güldemann (2018b:473)

L'utilité du concept d'aire linguistique a parfois été remise en question, par exemple par Campbell (2006), qui soutient qu'il s'agit d'une conception après coup qui n'est pas nécessaire en plus du concept d'emprunt. Cependant, l'héritage généalogique et la diffusion suite au contact n'épuisent en aucun cas les possibilités d'explication de la distribution géographique des traits typologiques. L'influence du substrat peut également jouer un rôle important dans l'explication des détails de la distribution actuelle d'un trait typologique et par conséquent qu'il est possible qu'un trait

typologique soit ancien dans la région où il se trouve tout en étant récent dans les familles linguistiques dans lesquelles il est actuellement réalisé. Ce principe est clairement illustré dans des exemples récents de relations substrat-superstrat, comme celle entre le portugais et les langues édoïdes dans les créoles du golfe de Guinée, qui sont des nouveaux arrivants dans la Macro-Sudan Belt tout en affichant nombre de ses caractéristiques typologiques (Güldemann & Hagemeijer 2015).

Un autre type d'explication diachronique du regroupement aréal des traits typologiques sous forme des aires linguistiques qui est rarement pris en compte dans la littérature, est que la présence d'un trait typologique peut être propice à l'émergence d'un autre. En d'autres termes, alors que les explications traditionnelles en termes d'héritage commun et de diffusion suite au contact ne peuvent voir dans le regroupement aréal des traits typologiques que des accidents historiques, on doit également tenir compte de la possibilité qu'il puisse y avoir de relations causales entre certains d'entre eux dans le cadre d'un modèle explicatif plus complexe. Ainsi, bien que je soutienne que la distribution géographique actuelle des fréquences lexicales des occlusives labiales-vélaires est due à l'influence de substrat (cf. §5.5), je reste agnostique quant à la présence ou non des occlusives labiales-vélaires dans les langues de substrat. L'explication de l'émergence, du maintien et de la diffusion des occlusives labiales-vélaires que je considère comme la plus probable est une explication en termes de causalité multiple qui implique la présence combinée d'un nombre de phénomènes phonétiques et phonologiques, dont le phénomène le plus important est la prosodie d'emphase-C (cf. §5.4). Deux autres phénomènes potentiellement pertinents que je ne discute pas ici sont la présence d'un contraste de position +/− avancée de la racine de la langue et de la réalisation endolabiale de certaines voyelle arrondies. Je propose également une explication similaire de causalité multiple pour la prépondérance des marques de négation en fin de phrase dans les langues de l'Afrique subsaharienne septentrionale (cf. §4). Ainsi, je soutiens qu'elle émerge à travers l'effet combiné d'un changement linguistique classique du cycle de Jespersen et d'une préférence générale dans les langues de l'Afrique subsaharienne septentrionale pour les marqueurs intersubjectifs tels que les intensificateurs à être en position finale d'énonciation (cf. Idiatov 2012a; 2015; 2018). Ce dernier type de trait est facilement emprunté (Matras 2009) et sa distribution aréale peut être expliquée en termes de diffusion suite au contact. De manière cruciale, l'explication de la distribution aréale de la négation en fin de phrase n'implique pas la diffusion suite au contact de ce trait lui-même.

3 Les objectifs et la méthodologie de la typologie aréale

3.1 Les objectifs

La typologie aréale a deux objectifs principaux qui sont logiquement hiérarchisés. Le premier objectif est la recherche de corrélations potentiellement intéressantes dans la distribution des valeurs de divers traits linguistiques dans l'espace. Ce premier objectif analytique pose les bases du deuxième objectif explicatif, qui est d'essayer de trouver des explications plausibles à la distribution observée. Les explications sont pour moi fondamentalement diachroniques, ou comme l'a formulé Dryer (2006:56) « a theory of why languages are the way they are is fundamentally a theory of language change ».

3.2 La méthodologie : les principes de base

Tout d'abord, je suis convaincu que du point de vue méthodologique c'est une bonne pratique de commencer par une hypothèse de recherche claire (formée de préférence par une connaissance approfondie des langues de la région), de définir l'hypothèse zéro et d'être conscient du biais possible qu'une décision particulière peut induire sur le résultat. L'approche méthodologique à la typologie aréale que j'ai développé dans mes recherches (cf. Idiatov 2018; Idiatov & Van de Velde 2021) a été conditionné par mon intérêt particulier au potentiel explicatif et historique de la typologie aréale et découle du constat que pour capter au mieux le signal diachronique dans les données synchroniques et arriver ainsi à une explication diachronique plus plausible, il faut tenir compte au maximum de toute la diversité dans les données disponibles.[3] De ce fait, je choisis une approche bottom-up pour laquelle le point de départ n'est pas des aires linguistiques prédéfinies, telles que Macro-Sudan Belt ou Balkan Sprachbund, que l'on essayerait de valider, mais des traits linguistiques et leur distribution dans l'espace. Pour capter au mieux le signal diachronique, aussi faible soit-il, je fais appel dans la mesure du possible à la méthodologie du big data dont la logique est que la quantité de données compenserait les éventuels problèmes de leur qualité ou les lacunes dans les données et

[3] Ceci va à l'encontre d'une autre approche méthodologique très courante qu'on pourrait qualifier de réductionniste. Ainsi, comme le constate Dryer (2009:316), « typological classification generally involves drawing arbitrary lines in what is really a typological continuum ». Toutefois, ce que l'approche réductionniste gagne en simplicité d'exécution, elle a tendance à perdre en pouvoir explicatif.

que la quantité révélerait les tendances générales des données. Il reste néanmoins claire que les données d'une meilleure qualité vont donner des résultats plus fiables suivant le principe analytique général *garbage in, garbage out*. Par conséquent, pour arriver à des résultats plus clairs, stables et surtout plus fiables, un effort important doit être investi dans le nettoyage des données et une attention particulière doit être prêtée aux valeurs aberrantes. Mon approche méthodologique de prédilection est également non binaire visant à utiliser des données quantitatives pour étudier non seulement si un trait est présent dans une langue, mais aussi dans quelle mesure il l'est, ou en d'autres termes, son degré d'enracinement, ce qui permet à émettre des hypothèses par exemple sur l'ancienneté de la présence du trait dans la langue ou sur la façon dont il y est entré. Une telle approche demande également l'utilisation des outils statistiques d'analyse des données qui permettent de préserver au maximum les détails fins des données évitant ainsi d'imposer des catégories arbitraires (faire le binning, en langage statistique) aux données qui sont par leur nature graduelles. Quand on passe à l'étape de l'interprétation qualitative des résultats et essaye de proposer des explications plausibles, les explications à préférer sont des explications en termes de scénarios qui impliquent des mécanismes concrets de changement linguistique, tels que l'innovation par changements de sons, l'héritage (la transmission verticale), l'emprunt des phonèmes par le biais de mots d'emprunt, l'interférence du substrat, et en évitant des concepts plus abstraits et moins clairement définis, tels que la diffusion.[4] De tels scénarios peuvent également être enrichis en utilisant des données provenant d'autres disciplines, tels que la géographie, l'ethnographie, l'histoire, l'archéologie, la paléobiologie, etc.

3.3 La méthodologie en pratique

3.3.1 La collecte, le nettoyage et la mise en forme des données

Du point de vue pratique, depuis une dizaine d'années l'application du côté big data de la méthodologie que je propose a été rendue plus facile par le développement des grandes bases de données en libre accès, dont la plus importante pour mes recherches est RefLex (Segerer & Flavier 2011–2022), qui recueille plus d'un millier de sources lexicales sur les langues africaines accompagnées d'un bon nombre d'outils d'exploitation. D'autres bases de données que j'ai utilisées sont Phoible (Moran, McCloy & Wright 2014; Moran & McCloy 2019), qui recueille les inventaires

[4] Du point de vue méthodologique, comparez l'argument de Van de Velde (to appear) pour l'approche en termes de scénarios (*scenario-based approach*) dans le domaine de la morphosyntaxe comparative sur l'exemple des langues bantu.

phonologiques de plus de deux milles langues, et GeoNames (www.geonames.org), qui contient plus de 25 millions de noms géographiques avec leur coordonnées géographiques. Bien que les données de telles bases de données puissent être moissonnées relativement facilement, un effort supplémentaire important doit normalement être investi dans le nettoyage des données et dans leur mise en forme en fonction des questions et hypothèses de recherche spécifiques et des présuppositions théoriques plus générales.

3.3.2 L'analyse spatiale des données et leur visualisation

Étant donné que la typologie aréale concerne l'analyse des distributions des traits typologiques dans l'espace, il est particulièrement important d'essayer de visualiser les données avec différentes méthodes de visualisation pour s'assurer que les résultats sont solides du point de vue qualitatif. On peut toujours essayer de commencer par les méthodes de visualisation les plus simples, tels que les motifs de points, qui sont toutefois aussi les moins fiables (§3.3.2.1). Dans cette perspective, je préfère commencer par des méthodes de visualisation déterministes (c.-à-d. non statistiques), tels que les graphiques d'interpolation spatiale (§3.3.2.2), qui ne sont pas particulièrement plus compliquées à réaliser et qui généralement produisent des résultats beaucoup plus fiables. Finalement, un niveau de fiabilité encore supérieur aux méthodes déterministes peut être obtenu avec des outils statistiques, tels que les modèles additifs généralisés (GAM) visualisés sous forme de tracés de contours (§3.3.2.3), qui permettent d'analyser des grandes masses des données tout en préservant au maximum les détails fins et en évitant d'imposer aux données des catégories arbitraires.

3.3.2.1 Les motifs de points

La méthode de visualisation la plus simple à réaliser est sans doute la visualisation à l'aide des motifs de points où les points de données sont visualisés sur la carte avec des couleurs différentes en fonction de leurs valeurs. Normalement, pour des fins de visualisation la plage des valeurs est divisée en un nombre de catégories (ou intervalles) limité, ce qu'en anglais on appelle *binning*. Inévitablement, la partition d'une plage de valeurs graduelles en catégories discrètes est fondamentalement arbitraire et par conséquent non objective, ce qui représente le plus grand défaut de cette méthode de visualisation.

À titre d'exemple, la figure 2 visualise à l'aide d'un motif de points un trait linguistique particulier dans un échantillon de 31 langues mande. Dans cet échantillon, la plage des valeurs du trait est de 0,93 à 3,49. On a décidé de diviser l'échantillon ordonnée par les valeurs du trait en trois groupes de taille égale, à savoir un groupe de 10 langues qui contient les valeurs de 0,93 jusqu'à 1,65, un autre groupe de 10 langues qui contient les valeurs de 1,65 jusqu'à 2,09 et un groupe de 11 langues qui contient les valeurs de 2,09 jusqu'à 3,49. Pour des raisons de référence, j'appellerai cette méthode de partition de l'échantillon en catégories comme *linéaire*.[5]

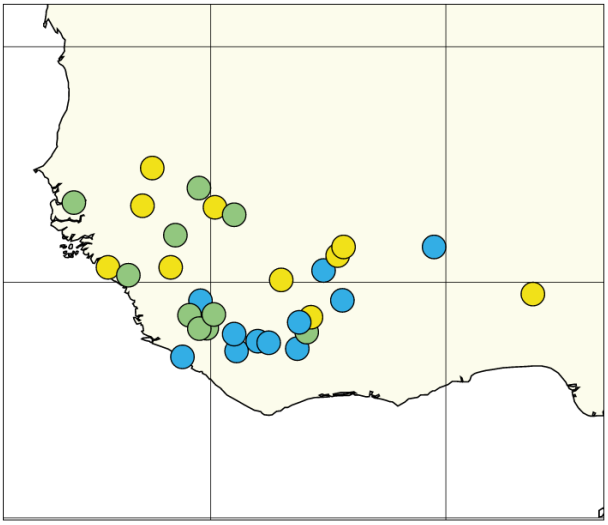


FIGURE 2. La visualisation d'un trait linguistique particulier dans un échantillon de 31 langues mande à l'aide d'un motif de points selon la méthode linéaire, à savoir en divisant l'échantillon en trois groupes de taille équivalente : ● groupe 1 (10 langues ; les valeurs : $0{,}93 \leq x < 1{,}65$) ; ● groupe 2 (10 langues ; les valeurs : $1{,}65 \leq x < 2{,}09$) ; ● groupe 3 (11 langues ; les valeurs : $2{,}09 \leq x \leq 3{,}49$).

Un coup d'œil sur cette visualisation suggère plusieurs motifs dans la distribution spatiale des valeurs de ce trait au sein de notre échantillon des langues mande. Ainsi, du point de vue purement géographique, les points bleus se trouvent surtout dans l'extrême sud et les points jaunes plutôt vers le nord, tandis que les points verts tendent plutôt vers la partie nord-ouest. En même temps, si l'on prend en compte également la sous-division généalogique des langues mande de cet échantillon, on peut remarquer que les points bleues sont surtout des langues mande sud-est. Les points verts sont surtout des langues mande ouest. Les points jaunes sont surtout des langues mande ouest également, mais ils contiennent aussi quelques langues mande est, qui est un sous-groupe des langues mande sud-est.

[5] Je repris ce terme *linéaire* (de l'anglais *linear*) de l'outil *Language Distribution* (https://reflex.cnrs.fr/core/Outils/Carto/index2.php) fourni par RefLex et que j'ai utilisé ici pour produire les visualisations à l'aide des motifs de points.

Ce simple exemple illustre bien l'un des problèmes principaux de cette méthode de visualisation, à savoir qu'il est difficile d'interpréter une telle visualisation de manière non ambiguë et surtout non subjective, parce que beaucoup dépend de la force d'imagination de l'observateur et de ses capacités de discerner des motifs dans la distribution des points des couleurs différentes dans l'espace. Un autre problème potentiel caractéristique de toutes les méthodes déterministes que je ne vais pas illustrer ici, est que les résultats de ce type de visualisation peuvent être particulièrement sensibles aux variations dans la qualité et la quantité des données. Par contre, je veux illustrer d'avantage le problème plus fondamental qui est particulier à cette méthode de visualisation et qui est dû à l'arbitraire de la partition de la plage des valeurs en intervalles.

Je commence avec un exemple moins frappant en utilisant le même échantillon de 31 langues mande. Au lieu de diviser les valeurs ordonnées en trois groupes de taille égale, on peut également décider de diviser la plage des valeurs de l'échantillon qui va de 0,93 à 3,49 en trois parties de taille égale, à savoir trois intervalles d'à peu près 0,85 de largeur. Ceci va donner un groupe de 14 langues qui contient les valeurs de 0,93 jusqu'à 1,78, un autre groupe de 11 langues qui contient les valeurs de 1,78 jusqu'à 2,63 et un groupe de 6 langues qui contient les valeurs de 2,63 jusqu'à 3,49. Pour des raisons de référence, j'appellerai cette méthode de partition de l'échantillon en catégories comme *proportionnelle*.[6] Le résultat est visualisé dans la figure 3 à l'aide d'un motif de points. Pour faciliter la comparaison, je reproduis d'abord également la figure 2 qui

[6] Je repris ce terme *proportionnel* (de l'anglais *proportional*) de l'outil *Language Distribution* (https://reflex.cnrs.fr/core/Outils/Carto/index2.php) fourni par RefLex.

visualise le même échantillon divisé en trois groupes de taille équivalente selon la méthode linéaire.

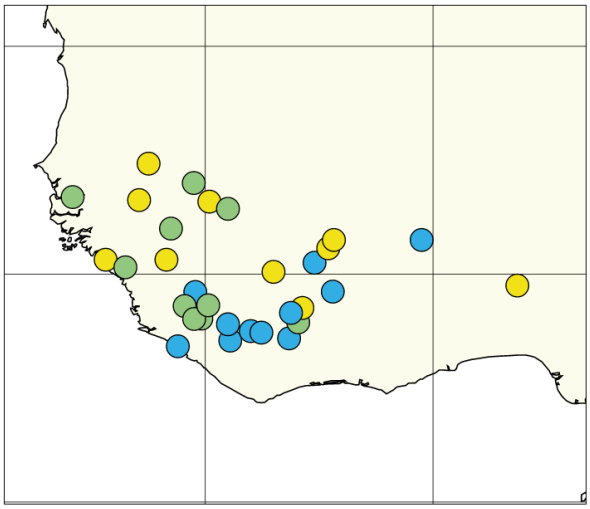


FIGURE 2. La visualisation d'un trait linguistique particulier dans un échantillon de 31 langues mande à l'aide d'un motif de points selon la méthode linéaire : ● groupe 1 (10 langues ; les valeurs : 0,93 ≤ x < 1,65) ; ● groupe 2 (10 langues ; les valeurs : 1,65 ≤ x < 2,09) ; ● groupe 3 (11 langues ; les valeurs : 2.09 < x < 3.49).

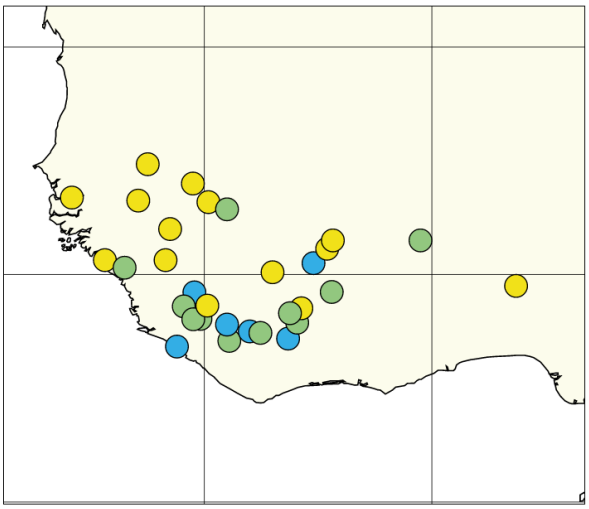


FIGURE 3. La visualisation d'un trait linguistique particulier dans un échantillon de 31 langues mande à l'aide d'un motif de points selon la méthode proportionnelle : ● groupe 1 (14 langues ; les valeurs : 0,93 ≤ x < 1,78) ; ● groupe 2 (11 langues ; les valeurs : 1,78 ≤ x < 2,63) ; ● groupe 3 (6 langues ; les valeurs : 2.63 < x < 3.49).

La méthode proportionnelle semble mettre plus en évidence l'aspect géographique dans la distribution spatiale des valeurs selon l'inclinaison nord-sud en confirmant également que les valeurs les plus extrêmes du trait représentées par les points bleues se trouvent surtout parmi les langues mande sud-est. Bien que les résultats de ces deux méthodes restent largement comparables, on n'a aucun moyen objectif de choisir entre les deux méthodes. En fin de compte, on pourrait également essayer de diviser notre échantillon non pas en trois catégories, mais en deux, quatre, ou cinq, ou même plus, ou essayer encore une autre méthode de division de l'échantillon en catégories.

Pour donner un exemple plus frappant des problèmes inhérents au caractère arbitraire de la partition d'une plage de valeurs graduelles en catégories discrètes, nous allons appliquer les mêmes deux méthodes de visualisation à un échantillon de 273 langues de l'Afrique subsaharienne pour le même trait que dans notre échantillon de 31 langues mande. Dans cet échantillon de 273 langues, la plage des valeurs du trait est de 0,43 à 10,49. La figure 4 visualise à l'aide d'un motif de points les résultats de la division de l'échantillon en trois catégories selon la méthode linéaire et la figure 5 visualise à l'aide d'un motif de points les résultats de la division de l'échantillon en trois catégories selon la méthode proportionnelle.

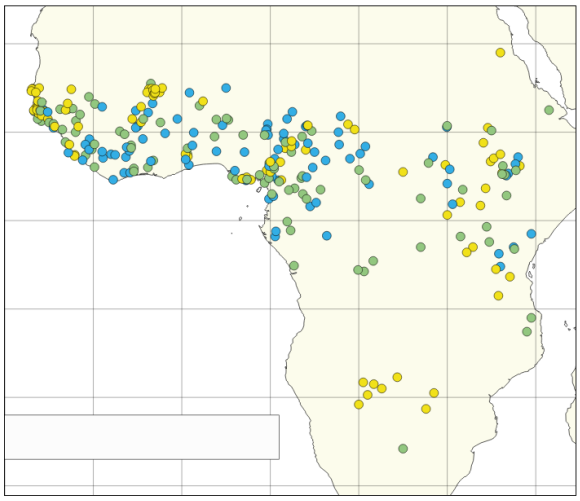


FIGURE 4. La visualisation d'un trait linguistique particulier dans un échantillon de 273 langues de l'Afrique subsaharienne à l'aide d'un motif de points selon la méthode linéaire : ● groupe 1 (91 langues ; les valeurs : 0,43 ≤ x < 1,36) ; ● groupe 2 (91 langues ; les valeurs : 1,36 ≤ x < 1,86) ; ● groupe 3 (91 langues ; les valeurs : 1,86 ≤ x ≤ 10,49).

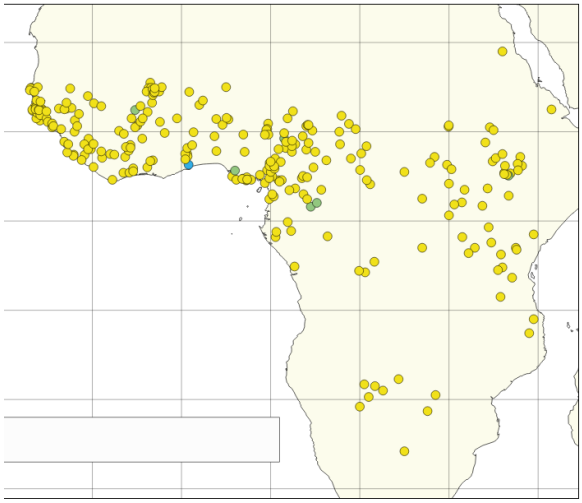


FIGURE 5. La visualisation d'un trait linguistique particulier dans un échantillon de 273 langues de l'Afrique subsaharienne à l'aide d'un motif de points selon la méthode proportionnelle : ● groupe 1 (266 langues ; les valeurs : 0,43 ≤ x < 3,78) ; ● groupe 2 (6 langues ; les valeurs : 3,78 ≤ x < 7,13) ; ● groupe 3 (1 langues ; les valeurs : 7,13 ≤ x ≤ 10,49).

Dans la visualisation selon la méthode linéaire, les points bleus semblent être limités à l'Afrique subsaharienne septentrionale, mais plus ou moins la même distribution semble également caractériser les points verts, tandis que les points jaunes semblent être plus fréquents sur les périphéries et le long de la frontière entre le Cameroun et le Nigeria et être absents de la côte du golfe de Guinée. Bien que la visualisation selon la méthode linéaire dans la figure 4 soit déjà compliquée à interpréter, la visualisation selon la méthode proportionnelle dans la figure 5 rend toute interprétation simplement inutile puisqu'elle groupe presque toutes les langues de l'échantillon dans une seule catégorie. La raison de ce déséquilibre est la présence de quelques valeurs extrêmes dans la partie supérieure de la plage des valeurs, à savoir les groupes 2 et 3 représentés par les points verts et bleus, tandis que la majorité des valeurs se concentrent dans la partie inférieure de la plage des valeurs, à savoir le groupe 1 représenté par les points jeunes.

3.3.2.2 Les méthodes déterministes : l'interpolation spatiale

Les méthodes déterministes, telles que certaines des méthodes d'interpolation spatiale couramment utilisées, produisent des modèles déterministes dont les résultats sont entièrement déterminés par les valeurs d'entrée et les fonctions mathématiques utilisées.

L’interpolation spatiale est un outil permettant de visualiser la distribution d’une variable dans l’espace en estimant la valeur d’une variable à un endroit spécifique sur la base d’une moyenne pondérée des valeurs connues à des endroits échantillonnés. Il existe un large éventail de méthodes d’interpolation spatiale. Pour mes recherches, j’applique deux méthodes déterministes d’interpolation spatiale couramment utilisées, la pondération inverse à la distance ou PID (en anglais, *inverse distance weighting* ou *IDW*) et le lissage par noyau (en anglais, *kernel smoothing*). Les deux méthodes donnent aux points les plus proches une pondération plus élevée, mais la méthode PID prend en compte tous les points de données connus, tandis que le lissage par noyau ne prend en compte que les points de données observés voisins dans une certaine fenêtre. Une autre différence importante entre les deux méthodes est que pour les endroits échantillonnés, la méthode PID produit des valeurs identiques aux valeurs observées (il s’agit d’un interpolateur exact), tandis que le lissage par noyau est comparable à la régression en ce sens qu’il peut produire des valeurs différentes des valeurs observées (il s’agit d’un interpolateur inexact). Par conséquent, le lissage par noyau est plus apte à mettre en évidence les tendances générales de la structure spatiale des données en lissant les pics et les creux aigus, tandis que la PID est plus apte à mettre en évidence les détails plus fins de la structure spatiale des données. La PID produit des graphiques un peu plus granuleux qui peuvent occasionnellement être perturbés par certains artefacts de visualisation trompeurs, car elle prend en compte tous les points de données connus.

Pour en donner un exemple, la figure 6 visualise un trait linguistique particulier dans un échantillon de 618 langues africaines à l’aide du lissage par noyau, tandis que la figure 7, visualise les mêmes données à l’aide de PID.[7] Les carrés vides dans les deux figures représentent la position dans l’espace des points de données (c’est-à-dire des langues de l’échantillon).

[7] Le trait linguistique en question est la négation en fin de phrase discuté dans §4. Pour une interprétation des deux figures du point de vue de la typologie aréale, voir §4.5.1.1.

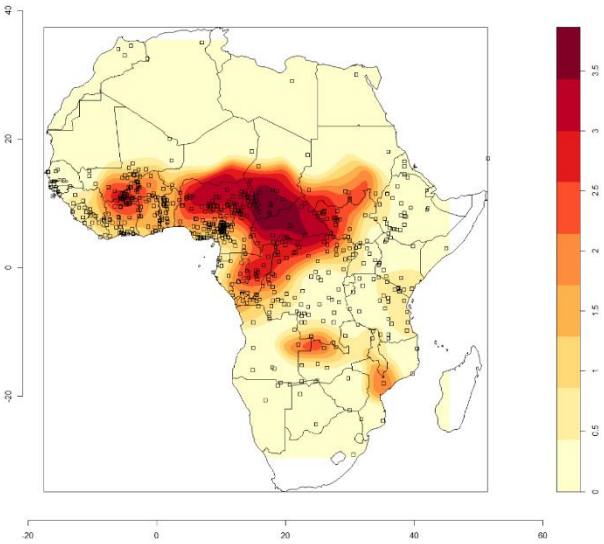


FIGURE 6. La visualisation par l'interpolation spatiale d'un trait linguistique particulier dans un échantillon de 618 langues à l'aide du lissage par noyau Gaussien (la valeur par défaut de la bande passante ajustée par 1,3).

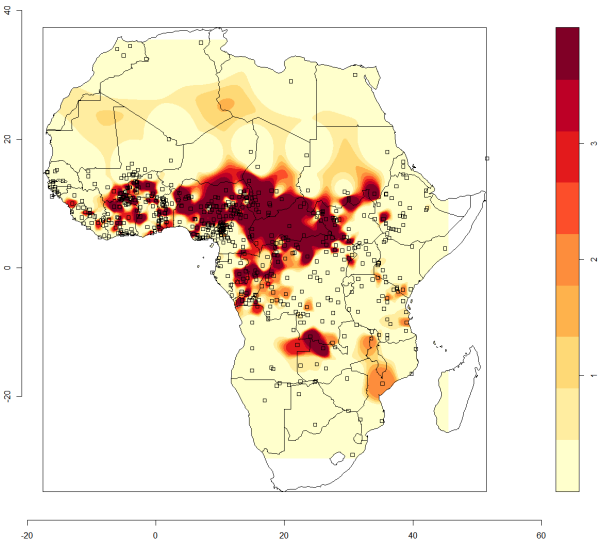


FIGURE 7. La visualisation par l'interpolation spatiale d'un trait linguistique particulier dans un échantillon de 618 langues à l'aide de la pondération inverse à la distance (la puissance = 6).

En fonction de la manière exacte dont les données sous-jacentes sont distribuées dans l'espace, l'interpolation spatiale peut produire certains artefacts de visualisation dont il faut être conscient lors de l'analyse des résultats. Ainsi, les deux méthodes visualisées dans les figures 6 et 7 exagèrent à des degrés différents la proéminence d'un certain nombre de régions, comme la région où les frontières de la République démocratique du Congo, de l'Angola et de la Zambie se rejoignent, la région au centre du Mozambique et la région vers le nord-est du Soudan du Sud. Ces régions à la proéminence exagérée sont dues au fait que les points de données de l'échantillon sont faiblement distribués dans ces régions (soit parce qu'il y a simplement moins de langues, soit parce que les langues n'ont pas été échantillonnées). Pour voir pourquoi cela peut affecter la visualisation, nous pouvons représenter les points de données par des pics dont la hauteur correspond à la valeur numérique du trait en question. Les pics représentant quelques exemples isolés de langues avec une valeur positive du trait dans une telle région auraient des pentes très larges lorsque les autres points de données sont éloignés. Plusieurs régions de faible proéminence au Sahara dans la figure 7 sont essentiellement dues à la même raison et ne correspondent à aucun point de données de l'échantillon. L'absence de ces fausses régions proéminentes au Sahara dans la figure 6 est simplement due à la façon dont les points de données de l'échantillon sont distribués sur les franges sud du Sahara et à la valeur de la largeur de bande passante choisie pour

le lissage du noyau.[8] Finalement, il convient d'être prudent dans l'interprétation de la région de haute proéminence (reflétée par son ombrage plus foncé) en Afrique centrale dans le sud du Tchad et de la République centrafricaine dans la figure 6. Bien qu'il soit tentant de l'interpréter comme le foyer historique de cette aire centrafricaine, une autre interprétation est également possible et dans ce cas particulier est en effet à préférer (cf. §4.5.2). À cet égard, il convient de noter que cette région très proéminente est absente de la figure 7, qui utilise une méthode d'interpolation spatiale différente.

Comme le montrent les légendes des figures 6 et 7, les différentes méthodes d'interpolation spatiale requièrent l'ajustement des paramètres de lissage supplémentaires qui influencent le résultat de la visualisation, tels que la bande passante pour le lissage par noyau et la puissance pour la PID. La bande passante spécifie la distance maximale à laquelle les points de données sont utilisés pour la prédiction. Lorsque la largeur de la bande passante augmente, le biais de prédiction augmente et la variance de prédiction diminue, ce qui résulte en moins de granularité dans la visualisation, ou en d'autres termes, en plus de lissage. Le paramètre puissance utilisé pour la PID contrôle l'importance des points environnants sur la valeur interpolée. Une puissance plus élevée entraîne une influence moindre des points éloignés et par conséquent plus de granularité dans la visualisation. Il est important de noter qu'il n'existe pas de méthodes formalisées qui seraient généralement acceptées pour identifier la « meilleure » valeur de ces paramètres de lissage. Ainsi, la comparaison et l'évaluation subjective de différentes valeurs des paramètres de lissage sur la visualisation est la méthode très généralement utilisées pour trouver le meilleur équilibre entre la préservation des détails les plus fins et le reflet des tendances générales. Ceci est particulièrement vrai pour le paramètre de puissance de la PID. Quant au paramètre de bande passante du lissage par noyau, il existe quand même un certain nombre de méthodes algorithmiques de son « optimisation », telles que le sélecteur de bande passante en deux étapes, l'erreur quadratique moyenne et la validation croisée, qu'on peut utiliser pour réduire la subjectivité dans le choix de la bande passante.

3.3.2.3 Les méthodes statistiques : les modèles additifs généralisés

Les modèles additifs généralisés (GAM) sont un outil statistique particulièrement bien adapté pour traiter des données graduelles complexes liées à des points dans l'espace. A l'origine, les GAMs sont une extension de la régression multiple qui fournit des outils flexibles pour modéliser des interactions complexes décrivant des surfaces ondulées.

[8] À cet égard, on peut remarquer également un certain nombre de légères pointes vers le nord dans la figure 6, comme les pointes dans le centre du Tchad et le sud-est du Niger, qui correspondent à des pointes beaucoup plus nettes aux mêmes endroits dans la figure 7.

Baayen (2013), Winter & Wieling (2016), Tamminga et al. (2016) et Wieling (2018) sont des introductions utiles à l'utilisation des GAMs en linguistique. Des exemples d'utilisation des GAMs en linguistique en relation avec l'analyse spatiale peuvent être trouvés dans Wieling et al. (2011; 2014). Outre leur capacité à traiter des données hautement non linéaires, l'un des grands avantages des GAMs est qu'ils constituent un outil permettant aux données complexes de parler d'elles-mêmes sans avoir à les recoder ou à les classer au préalable. Cependant, la liberté qu'offrent les GAMs et leur capacité à traiter des données très complexes ont également un certain effet secondaire. Ainsi, les coefficients qui accompagnent les GAMs ne peuvent pas être facilement interprétés de manière directe et leur visualisation est très importante pour leur évaluation.

Les GAMs peuvent être visualisés de plusieurs manières. Dans mes recherches, je trouve particulièrement pratique la visualisation des GAMs sous forme de tracé de contours avec le schéma de couleurs de la carte thermique. Cette visualisation représente la surface de régression de la variable dépendante, comme les estimations des fréquences lexicales des occlusives labiales-vélaires (cf. §5) ou le type du marquage de négation (cf. §4), en fonction de la combinaison de la longitude et de la latitude en utilisant des splines de régression à plaque mince (« thin-plate regression splines »). Les teintes claires correspondent à des valeurs plus élevées de la variable dépendante. Les lignes de contour sont des isoplèthes qui marquent les déviations de la moyenne en termes d'écart-type. Un paramètre supplémentaire qui est assez important pour une telle visualisation des GAMs est le paramètre qui contrôle la taille de la zone à tracer autour de chaque point de données.[9] Ce paramètre dont le réglage reste subjectif n'influence toutefois que la visualisation du modèle et non pas le modèle lui-même. En choisissant ce paramètre, j'essaye de trouver un équilibre entre la précision de la représentation de la continuité spatiale entre les points de données et la facilité de perception de la visualisation dans son ensemble, en évitant trop de discontinuités dans le tracé des contours. La figure 8 illustre la visualisation d'un GAM avec le schéma de couleurs de la carte thermique où

[9] Dans le paquet *mgcv* pour *R* (Wood 2006; 2019) que j'ai utilisé pour produire des GAMs et leurs visualisations, ce paramètre s'appelle *too.far* et sa valeur par défaut est 0,1.

le paramètre qui contrôle la taille de la zone à tracer autour de chaque point de données a été fixé à 0,5 (la moitié de sa valeur par défaut).

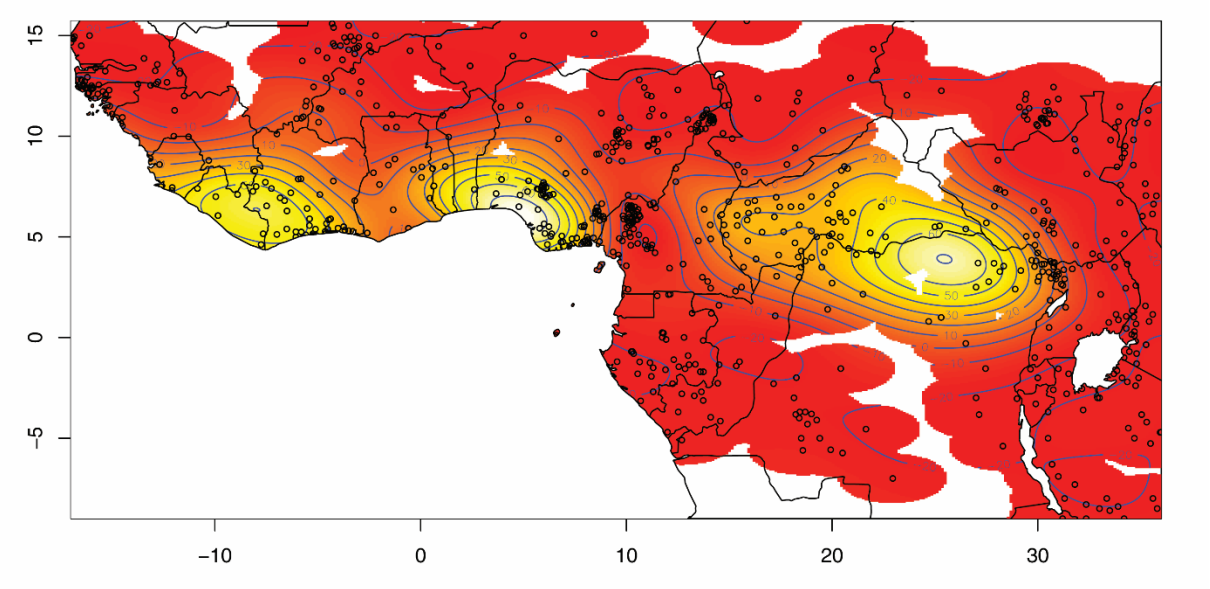


FIGURE 8. Le tracé de contours avec le schéma de couleurs de la carte thermique représentant la surface de régression GAM des fréquences lexicales des occlusives labiales-vélaires (F_{LV}) en pourcentage (y compris 0% pour les langues sans occlusives labiales-vélaires) en fonction de la combinaison de la longitude et de la latitude en utilisant des splines de régression à plaque mince. Le résumé du modèle : k = 13 (k-index = 0.99, p = 0.39, k′ = 195), fonction = gaussienne, edf = 70.76, la déviance expliquée = 77.60%, AIC = 6048, intercept F_{LV} = 19.95%, p < .001.

La principale différence conceptuelle entre la modélisation additive généralisée et les méthodes d'interpolation spatiale discutées dans §3.3.2.2 est que ces dernières méthodes produisent des modèles déterministes dont les résultats sont entièrement déterminés par les valeurs d'entrée et les fonctions mathématiques utilisées, tandis que la modélisation additive généralisée produit des modèles statistiques décrivant une distribution de résultats possibles. En termes pratiques par rapport à mes recherches dans la typologie aréale, la GAM produit généralement des visualisations plus claires qui, fait important, sont beaucoup plus stables aux variations de la couverture des données d'entrée ce qui par conséquent peut fournir une confirmation supplémentaire de la robustesse des résultats. J'en fournis une illustration dans §5.3.2.2 sur l'exemple des fréquences lexicales des occlusives labiales-vélaires dans le nord de l'Afrique subsaharienne en comparant les visualisations GAM basées sur l'ensemble complet de données et des visualisations GAM basées sur différents sous-ensembles des mêmes données.

Une autre différence pratique entre la GAM et les deux méthodes d'interpolation spatiale est que la GAM produit des résultats quantifiés ce qui est utile pour au moins deux raisons. Premièrement, les résultats quantifiés de la GAM peuvent nous aider à repérer des distributions intéressantes dans les données qui seraient beaucoup plus

difficiles à identifier autrement (cf. §5.3.2.3 pour un exemple). Deuxièmement, la GAM nous permet d'évaluer de manière plus objective la qualité de la visualisation et la précision du modèle statistique qui la produit. Il y a deux questions principales ici. La première est le niveau de précision des estimations des coefficients des modèles auquel on peut s'attendre de façon réaliste avec notre type de données. La deuxième question principale est de savoir si les modèles produits sont qualitativement robustes.

Le niveau de précision des estimations des coefficients d'un GAM basé sur la distribution gaussienne dépend de la mesure dans laquelle le modèle satisfait à la condition que les résidus sont normalement distribués et à la condition que la variance des résidus est constante (homoscédastique) pour toutes les valeurs du prédicteur linéaire. Pour certaines questions de recherche, comme les études sur l'importance de différences minuscules dans certains phénomènes phonétiques particulièrement fins, la précision des estimations des coefficients est très importante, car même un changement mineur dans la valeur des estimations des coefficients peut affecter la nature de nos inférences (par exemple, cf. Wieling 2018). Une précision quantitative élevée est beaucoup moins pertinente pour le type de données qu'on examine typiquement dans la typologie aréale, où beaucoup d'imprécision est intrinsèquement intégrée aux données, puisque typiquement la variable dépendante, comme les estimations des fréquences lexicales des occlusives labiales-vélaires, et la variable indépendante, à savoir la combinaison des valeurs de longitude et de latitude prises pour représenter la localisation des langues de l'échantillon, sont nécessairement des approximations grossières. Ce qui importe le plus, c'est la robustesse qualitative des résultats qui pourrait être confirmée d'avantage par une validation par recoupement utilisant différentes méthodes (interpolation spatiale et GAM), différents types de sous-échantillons (cf. §5.3.2.2) ou éventuellement différents types de données (cf. §5.3.3).

Bien que la GAM soit un bon outil pour décrire des surfaces ondulées, elle peut avoir des problèmes avec de grands changements abrupts dans la valeur de la variable dépendante alors que la valeur de la variable indépendante, c'est-à-dire la combinaison de la longitude et de la latitude, ne change que très peu. Par conséquent, des sauts ou des creux locaux abrupts dans les valeurs de la variable dépendante peuvent entraîner des valeurs aberrantes dans les résidus de la surface de régression produite par un GAM.

Ces fluctuations locales abruptes de la valeur dépendante que nous pouvons repérer si facilement en utilisant le graphique des résidus par rapport au prédicteur linéaire d'un GAM sont particulièrement intéressantes pour l'analyse des données pour deux raisons. Tout d'abord, elles peuvent mettre en évidence les sources de nos données qui peuvent nous donner des estimations moins précises de la valeur dépendante. Les estimations provenant de ces sources doivent être validées par recoupement avec de meilleures sources, si elles sont disponibles. Deuxièmement, outre ces imperfections

potentielles dans les données utilisées, les fluctuations locales abruptes de la valeur dépendante mises en évidence par l'inspection des résultats quantifiés du GAM peuvent également fournir une fenêtre sur la dynamique spatio-temporelle actuelle, car elles reflètent très probablement des événements historiques relativement peu profonds. Si de telles fluctuations abruptes s'étaient produites à des profondeurs temporelles plus significatives, nous nous attendrions à ce qu'elles aient été davantage lissées à l'heure actuelle. Deux types de processus spatio-temporels principaux peuvent se manifester, à savoir des événements récents de perte ou d'émergence locale du trait en question et des événements récents de propagation ou de relocalisation de langues qui auraient fait converger dans une petite région des langues présentant des profils de trait en question sensiblement différents (y compris l'absence totale du trait). Comme je l'explique dans §5.3.2.3 avec l'exemple des fréquences lexicales des occlusives labiales-vélaires, ces processus peuvent produire deux types de configurations dans la distribution spatiale de la valeur dépendante qui peuvent être conceptualisés comme des pics et des falaises positifs ou négatifs respectivement.

4 La négation en fin de phrase en Afrique

4.1 Introduction

Les études telles que Beyer (2009), Dryer (2009) et Devos & van der Auwera (2013) ont attiré l'attention sur certaines propriétés typologiquement frappantes du marquage de la négation dans les langues de différentes régions d'Afrique. Dryer (2009) se concentre sur la négation « neutre » (« neutral negation »), c'est-à-dire les modes obligatoires et productifs de marquage de la négation (globale) dans les phrases principales verbales déclaratives exprimés par des marques de négation qui sont des mots, dans les langues à ordre SVO en Afrique. Il démontre que les langues SVO dans une région d'Afrique centrale qui s'étend du Nigeria à la République centrafricaine et au nord de la République démocratique du Congo, comme l'illustre la figure 9, diffèrent considérablement des langues SVO du reste du monde en ce que « the negative [word] follows the verb [instead of preceding it], typically occurring at the end of the clause, in SVONeg order » (Dryer 2009:307). Dryer (2009) souligne également que le double marquage de la négation est très répandu dans cette région.

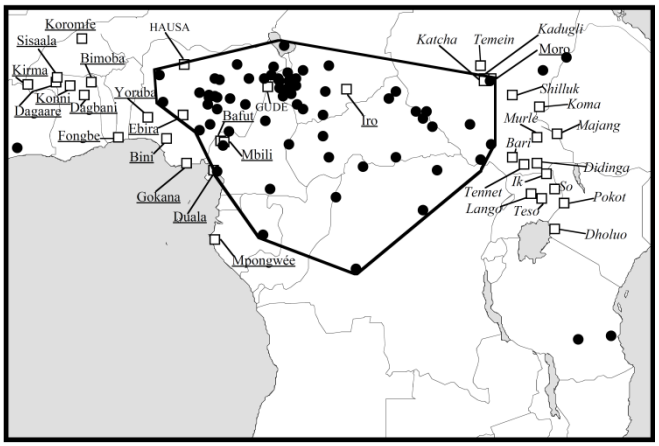


FIGURE 9. Les langues VO&VNeg en Afrique, avec leur aire principale délimitée (Dryer 2009:323)

Beyer (2009), dans le même volume sur les modes de négation dans les langues ouest-africaines (Cyffer, Ebermann & Ziegelmeyer 2009) que Dryer (2009), se concentre spécifiquement sur le double marquage de la négation pour la négation phrastique (« sentential negation ») dans un grand groupe de langues ouest-africaines centrées sur

le bassin de la Volta, comme illustré dans la figure 10. Dans la plupart des cas, la seconde des deux marques de négation se trouve également en fin de phrase.

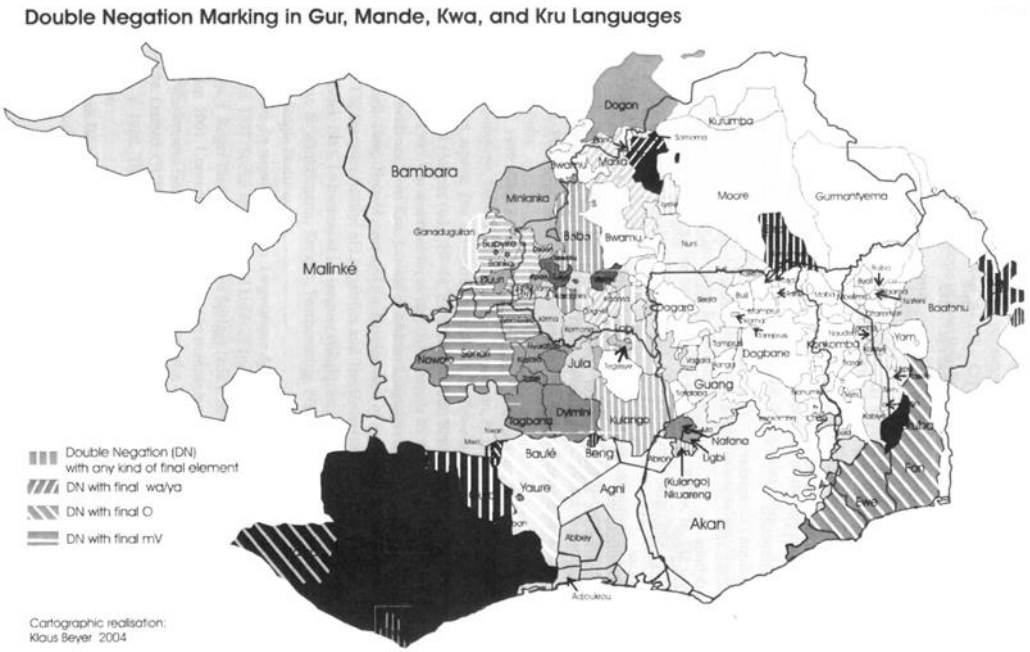


FIGURE 10. Le double marquage de la négation dans les langues ouest-africaines centrées sur le bassin de la Volta (Beyer 2009:222)

Devos & van der Auwera (2013) est une étude approfondie du marquage multiple de la négation dans les langues bantu. Ils étudient les cas de négation multiple dans les langues bantu, qui sont généralement doubles, mais certains exemples exubérants de triple et quadruple marquage de négation sont également attestés. Ils étudient également les sources récurrentes de marques de négation post-verbales. Beaucoup de ces marques de négation post-verbales se trouvent être également en fin de phrase, comme illustré dans la figure 11.[10]

[10] Dans la tradition bantuisante, le terme « post-final » utilisé dans la figure 11 fait référence à la position qui suit immédiatement le verbe.

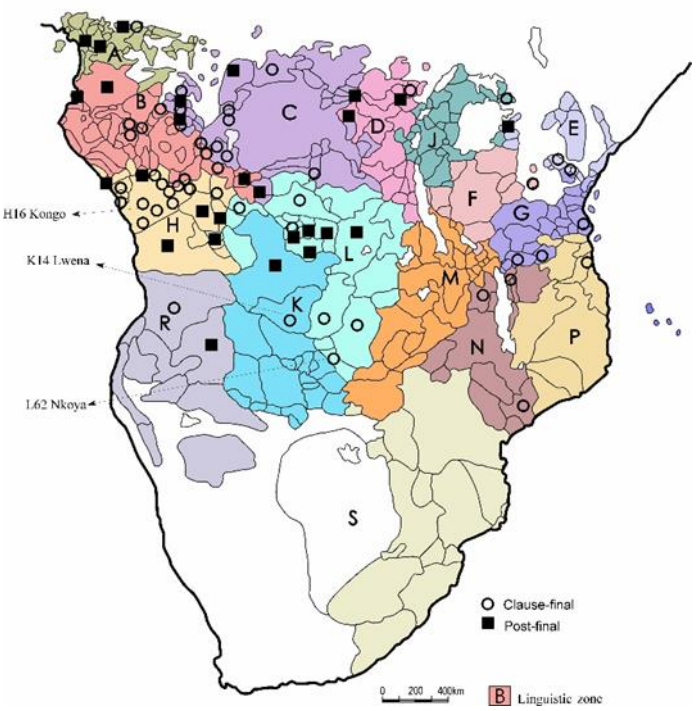


FIGURE 11. Double négation en bantu (Devos & van der Auwera 2013:215)

Comme ces études le montrent clairement, les marques de négation en fin de phrase (MNFPs), bien que typologiquement rares, peuvent être trouvées dans un très large éventail de langues d'Afrique subsaharienne. Sur la base d'un échantillon de 618 langues africaines, je démontre que la distribution spatiale des langues comportant des MNFPs forme un profil aréal clair en Afrique subsaharienne. Dans le même temps, la distribution spatiale des 462 langues comportant des marques de négation post-verbales de quelque nature que ce soit ne forme aucun profil aréal distinctif, car elle est pratiquement identique à la distribution spatiale de toutes les langues de l'échantillon dans son ensemble. Les deux distributions superposées avec leurs courbes d'intensité spatiale sont présentées respectivement dans la figure 12 et la figure 13.[11]

[11] Tous les graphiques et calculs ont été réalisés avec le programme R (R Core Team 2015) en utilisant RStudio IDE (RStudio Team 2016). J'ai utilisé le paquet *spatstat* pour produire les graphiques d'intensité spatiale et d'interpolation spatiale (Baddeley & Turner 2005).

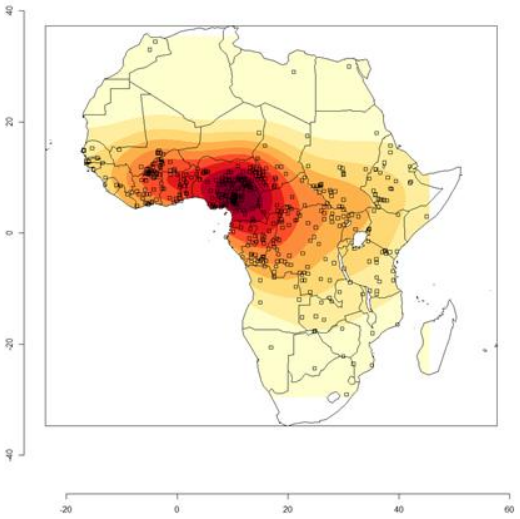


FIGURE 12. La distribution géographique des 462 langues de l'échantillon avec des marques de négation post-verbales et un graphique de leur intensité spatiale.

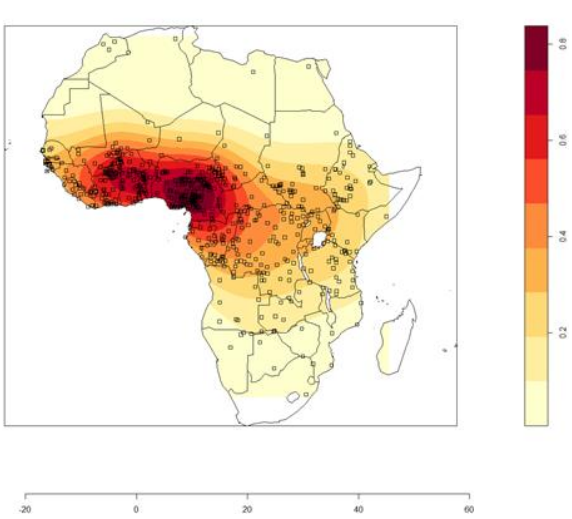


FIGURE 13. La distribution géographique des 618 langues de l'échantillon et un graphique de leur intensité spatiale.

En plus de former un profil aréal clair au sein de l'Afrique subsaharienne, à l'échelle mondiale les MNFPs sont aussi typologiquement beaucoup plus inhabituelles que les marques de négation post-verbales et le marquage multiple de la négation. En outre, comme je l'ai soutenu ailleurs (Idiatov 2012a), les MNFPs en Afrique subsaharienne ont tendance à être caractérisées par un certain nombre de particularités dans leur morphosyntaxe et leur développement diachronique qui les distinguent des marques similaires ailleurs dans le monde et offrent des indices importants quant à une explication de leur distribution aréale observée. Bien sûr, certaines de ces différences sont plus une question de degré, mais certaines semblent être plus fondamentales. Par exemple, les MNFPs dans les langues africaines sont souvent associés à la présence du marquage multiple de la négation dans une phrase, le plus souvent double mais aussi parfois triple et occasionnellement quadruple. Les MNFPs en Afrique sont souvent morphosyntaxiquement déficientes par rapport aux marques grammaticales plus canoniques, car elles sont optionnelles ou absentes dans certains types de phrases, selon la valeur TAM du prédicat de la phrase, le statut de subordination de la phrase, l'information structurelle associée et les valeurs du type d'acte de parole ou le type de discours auquel la phrase appartient (cf. Idiatov 2015). Diachroniquement, les MNFPs de la région ont tendance à être plutôt instables et semblent pouvoir être empruntées relativement facilement (cf. Idiatov 2012b; 2015), contrairement aux marques de négation dans d'autres parties du monde mais plutôt comme des marqueurs de discours, des particules de focalisation et des adverbes phasiques (cf. Matras 2009).

Tout cela fait des MNFPs en Afrique subsaharienne un trait morphosyntaxique particulièrement intéressant à explorer du point de vue de la dynamique des langues dans

l'espace et dans le temps. Ce chapitre propose une telle analyse spatio-temporelle du trait MNFP Pour des raisons d'espace, je ne m'étendrai pas dans le reste du chapitre sur l'explication de la raison pour laquelle de nombreux marques de négation sont finales dans la phrase en Afrique subsaharienne et pourquoi ces marques de négation sont si communes dans cette région en particulier. Je traite ces questions plus en détail ailleurs (Idiatov 2012a; in prep.). Je me limite ici donc à esquisser l'essentiel de l'explication, qui se présente comme suit. La position en fin de phrase des marques de négation s'explique par leur origine dans d'autres marques de fin de phrase. Le fait que les MNFPs soient si communes dans cette région est lié à une autre caractéristique typologique de beaucoup des langues concernées, à savoir une catégorie grammaticale de marques de fin de phrase dont la fonction principale est l'expression de fonctions intersubjectives. Si l'on ajoute à cela le fait que la négation est exactement l'une de ces situations, propice à l'utilisation de marques intersubjectives, où l'autorité affirmée du locuteur est en jeu, les effets de fréquence expliquent naturellement la tendance à conventionnaliser les marques de négation en fin de phrase.

Le reste de ce chapitre est organisé comme suit. Je commence par discuter dans §4.2 de divers aspects de la définition des MNFPs adoptée pour cette typologie. Cette définition est plutôt inclusive car, comme je l'explique dans §4.2.1, mon objectif est de capturer la plus grande partie de la diversité synchronique pour obtenir une meilleure adéquation explicative. Cependant, pour les raisons expliquées dans §4.2.2, je laisse de côté les constructions de négation des prédicats nominaux. Puisque je considère à la fois les MNFPs obligatoires et les MNFPs facultatives, dans §4.2.3 j'aborde certains des problèmes que la distinction entre les deux peut présenter. Dans §4.2.4, je développe la signification du terme « fin de phrase » dans cette typologie. La question de la pertinence de l'ordre relatif de l'objet et du verbe pour une typologie des MNFPs est liée à ce dernier point. Comme je l'explique dans §4.2.5, cet ordre n'est pas pertinent dans la typologie présentée ici, contrairement par exemple à la typologie des marques de négation post-verbales de Dryer (2009), qui se limite aux langues avec un ordre VO. Je présente brièvement mon échantillon dans §4.3. Je fournis également des cartes des 618 langues de l'échantillon dans son ensemble et des langues avec et sans MNFPs. La simple division binaire entre les langues qui ont des MNFPs et celles qui n'en ont pas cache une diversité importante parmi les langues avec MNFPs. Comme je l'explique dans §4.4, afin de mieux saisir cette diversité et d'avoir ainsi une meilleure idée de la dynamique historique et spatiale possible derrière la distribution observée, j'augmente le degré de granularité de mes données en prenant en compte deux paramètres, à savoir le caractère obligatoire des MNFPs et les éventuelles restrictions à la liberté d'utiliser les MNFPs dans différentes constructions. §4.5 présente une discussion de la dynamique spatiale et temporelle reflétée dans les profils typologiques aréaux observés des MNFPs

en Afrique. Dans §4.5.1, je discute d'abord des résultats et des pièges potentiels de deux méthodes d'analyse spatiale et de visualisation de la distribution des valeurs de la caractéristique MNFP en Afrique, à savoir l'interpolation spatiale et la modélisation additive généralisée (GAM). Les deux méthodes convergent sur la nécessité de distinguer deux zones de convergence du trait MNFP. La première, l'Aire de Convergence Centrale, est la plus importante des deux et s'étend à l'est de l'Afrique de l'Ouest et à certaines parties de l'Afrique centrale, coïncidant largement avec la zone centrale des langues VO&VNeg de Dryer (2009) reproduite dans la figure 9. La seconde, l'Aire de Convergence Occidentale, est moins importante et se limite à l'Afrique de l'Ouest. Les deux aires de convergence sont séparées par une discontinuité majeure autour du Ghana, du Togo et du Bénin. Dans §4.5.2, je fais appel à d'autres types de données pour mieux calibrer les résultats de l'analyse spatiale produite dans §4.5.1 et pour identifier le foyer historique de l'Aire de Convergence Centrale. Enfin, §4.5.3 traite de la distribution des MFNPs facultatives et/ou restreintes en Afrique, avec un accent particulier sur la diffusion des MNFPs parmi les langues bantu au sud de l'Aire de Convergence Centrale, principalement dans le corridor du fleuve Congo et au nord de la République démocratique du Congo.

4.2 Quels types de MNFPs est-ce qu'on prend en considération ?

4.2.1 Une définition inclusive : la diversité synchronique comme fenêtre sur le changement linguistique

Les propriétés morphosyntaxiques des constructions négatives diffèrent selon les langues. De même, le marquage de la négation peut varier dans une langue donnée d'une construction prédicative à une autre. Il existe de nombreux paramètres différents selon lesquels la variation se produit. En fonction de nos objectifs et de nos moyens, nous pouvons découper cet espace de variation de différentes manières. La définition que j'adopte ici est plutôt inclusive puisque mon objectif est de capturer la plus grande partie de la diversité. Le raisonnement qui sous-tend cette définition est que la diversité synchronique reflète directement la nature graduelle du changement linguistique et nous offre ainsi une fenêtre sur les processus historiques qui ont conduit à la situation actuelle. Ainsi, l'un des mécanismes les plus courants connus pour être impliqués dans l'évolution des constructions négatives, le cycle dit de Jespersen (cf. van der Auwera 2009; 2010 pour une vue d'ensemble; Devos & van der Auwera 2013 sur les langues bantu), passe par un certain nombre d'étapes avec la plupart des étapes intermédiaires

étant caractérisées par une variation du marquage de la négation dans une construction donnée. Typiquement, les langues apparentées ne suivent pas ce chemin exactement de la même manière. Tant la variation synchronique au sein d'une langue que la diversité synchronique des schémas de négation au sein d'un groupe de langues apparentées offrent une source inestimable d'informations sur les stades antérieurs des langues respectives et sur les processus qui sous-tendent le changement des constructions négatives.

Dans le cadre de la présente typologie, je considère comme MNFPs les éléments qui peuvent être utilisés dans la périphérie droite des prédications verbales négatives avec négation à portée de phrase mais qui n'apparaissent pas dans les prédications positives correspondantes et dont la position est déterminée par rapport à la phrase dans son ensemble. Une MNFP peut être la seule marque de négation dans la phrase ou juste un des exposants du marquage de négation distribué dans la phrase. Autrement dit, ma typologie ne se limite pas au double marquage de la négation, comme la typologie de Beyer (2009). Une MNFP peut être une marque de négation dédiée ou peut également encoder d'autres sens, comme le temps, l'aspect, le mode et l'emphase (à cet égard, voir également §4.2.3). Dans le cadre de ma typologie, le degré d'attachement morphologique des MNFPs n'est pas non plus pertinent. En d'autres termes, elles peuvent être des mots (comme les marques de négation dans la typologie de Dryer 2009), des clitiques, des affixes (supra)segmentales ou des opérations morphologiques non linéaires. De même, ma typologie prend en considération la négation de tous les types de phrases verbales et ne se limite pas à la négation des phrases principales verbales déclaratives, comme la typologie de la « négation standard » de Miestamo (2008) ou la typologie des « neutral clausal negatives » de Dryer (2009). La négation des prédicats nominaux est en dehors du cadre de ma typologie pour les raisons exposées dans §4.2.2 Comme discuté dans §4.2.3, je considère à la fois les MNFPs obligatoires et facultatives. La signification de la description « en fin de phrase » dans les MNFPs est expliquée plus en détail dans §4.2.4. §4.2.5 traite plus en détail du concept de fin de phrase par rapport aux marques de négation du point de vue des différents ordres relatifs de l'objet et du verbe.

4.2.2 En dehors de la considération de la typologie : la négation des prédicats nominaux

Je ne m'intéresse pas ici aux constructions de négation avec des prédicats nominaux. La raison n'est pas que je ne les considère pas comme pertinentes. Il est clair qu'il est important de prendre également en compte les stratégies de négation utilisées avec des prédicats nominaux si l'on veut obtenir un compte rendu diachronique complet des

constructions de négation phrastique avec des prédicats verbaux. Ainsi, comme l'a souligné Croft (1991), les marqueurs existentiels négatifs peuvent être étendus aux prédicats verbaux négatifs au sein de ce que l'on appelle le « negative-existential cycle ». Cependant, du point de vue d'une typologie aréale des MNFPs, les constructions de négation avec prédicats nominaux tendent à présenter des types de problèmes analytiques assez différents. Par exemple, dans le cas des constructions négatives existentielles (à distinguer des locatives-présentatives ; cf. Veselinova 2013) qui utilisent une marque de négation dédiée sans aucune marque existentielle distincte, la question de savoir si la position de cette marque est déterminée par rapport à la phrase dans son ensemble ou par rapport au prédicat nominal peut être simplement non pertinente. Pour les constructions équationnelles et d'identification, il n'est pas toujours évident de savoir quel nominal doit être considéré comme le prédicat (cf. Bisang & Sonaiya 2000 sur les constructions Yoruba de la structure X 'ÊTRE' Y, où X et Y sont des nominaux). Compte tenu de ces complications, une autre raison pour laquelle j'ai décidé de ne pas inclure la négation des prédicats nominaux dans la typologie aréale des MNFPs est que j'ai la forte impression que leur inclusion n'affecterait pas de manière significative le profil aréal établi uniquement sur la base des constructions avec prédicats verbaux. Ainsi, je n'ai trouvé que quelques langues décrites avec des MNFPs seulement dans les constructions négatives avec des prédicats nominaux mais pas dans celles avec des prédicats verbaux, comme le ngangela [nyem1238] (Bantu K12 ; Maniacky 2003) (Bantu K12 ; Maniacky 2003), qui a une MNFP optionnel *ko* seulement avec des prédicats nominaux comme illustré dans (1) versus (2), et le beiya [gomn1238] (Samba Duru ; Littig & Kleinewillinghöfer 2012), qui possède une MNFP *ʔwə́* uniquement avec des prédicats nominaux comme illustré en (3) versus (4) (pour ces deux langues, les sources ne fournissent que des exemples de constructions d'identification). Bien que je ne considère pas ces langues comme ayant une MNFP aux fins de la typologie présentée ici, leur ajout ne perturberait pas le profil aréal général établi sur la base des constructions de négation avec des prédicats verbaux uniquement.

ngangela [nyem1238]

(1) a. ***kací*** *impweevó* ***ko***
NEG.COP femme NEG

b. ***kaci*** *ímpweevo*
NEG.COP femme

c. ***kéci-ko*** *ímpweevo*
NEG.COP-NEG femme
'Ce n'est pas une femme.' (Maniacky 2003:192)

(2) ***ko****-tw-a-mween-****e*** *ðiŋgóómbe*
NEG-1PL-PRF-voir.PRF-NEG vaches
'Nous n'avons pas vu les vaches.' (Maniacky 2003:140)

beiya [gomn1238]

(3) *yɔ́ɔ́* *yēn* *kúsén* ***ʔwɔ́***
COP chose brousse NEG
'Ce n'est pas un animal sauvage.' (Littig & Kleinewillinghöfer 2012:6)

(4) *miǹ* *túúrɔ́* *Fɔ̀lé*
1SG venir**NEG** PROP
'Je ne viens pas à Poli.' (Littig & Kleinewillinghöfer 2012:6)

4.2.3 La question de l'optionalité

Je considère à la fois les MNFPs obligatoires et optionnelles. Les MNFPs obligatoires peuvent être obligatoires dans toutes les constructions négatives ou être limitées à un sous-ensemble de celles-ci. Les éléments optionnels des constructions négatives sont pris en considération dans la mesure où leur ajout ne change pas le sens propositionnel de la prédication négative ou que les contraintes sur leur utilisation sont conditionnées principalement par des propriétés structurelles de leur environnement plutôt que par leur sens (cf. Idiatov 2015 sur la MNFP *wāā* dans la langue mande Dzuun [dzuu1241]). Certes, il n'est pas toujours possible de faire une distinction nette selon ces lignes, précisément parce que le changement de langue est graduel. L'un des cas fréquents de ce genre est représenté par des éléments que l'on dit être ajoutés facultativement pour « accentuer » la négation et qui sont parfois pourvus de traductions telles que 'du tout'. En règle générale, je présume que si l'auteur d'une description grammaticale juge nécessaire d'indiquer qu'une construction de négation peut contenir un élément

optionnel donné, cet élément est suffisamment fréquent dans cette construction pour que sa signification référentielle originale soit suffisamment mise en arrière-plan.

Un autre type de situation qui est souvent conçu comme impliquant l'optionalité est celui où une marque de négation par défaut peut être remplacée par une marque de négation qui change le sens propositionnel de la prédication négative et, pour cette raison, aucune des deux marques ne peut être considéré comme obligatoire en tant que tel, mais la présence d'au moins une telle marque dans la construction est requise pour que la construction soit négative. En d'autres termes, c'est la manière particulière d'exprimer la négation dans une construction de négation qui est obligatoire, mais pas les marques de négation spécifiques. Le français fournit un bon exemple d'une telle situation suite à une évolution continue de type cycle de Jespersen et de la perte de la concordance négative (cf. van der Auwera & Van Alsenoy 2016). En français familier, l'ancienne marque de négation préverbal *ne* est généralement omise et seul la nouvelle marque de négation *pas* est utilisée immédiatement après le verbe, comme dans (5). La marque de négation par défaut pas peut être remplacée par un certain nombre de marques plus spécifiques, tels que *jamais*, comme dans (6), ou *nulle part*, comme dans (7), et, bien que ces derniers éléments changent le sens propositionnel de la prédication, au moins un élément de ce type doit être utilisé dans cette position dans la construction pour que la prédication reste négative.

français

(5) *Elle* ***(ne)*** *va* ***pas****.*

(6) *Elle* ***(ne)*** *va* ***jamais****.*

(7) *Elle* ***(ne)*** *va* ***nulle part****.*

Un exemple un peu plus compliqué est fourni par la langue mande dzuun [dzuu1241], comme discuté par Idiatov (2015). Ainsi, le dzuun possède une MNFP par défaut *wāā*, comme dans (8), qui peut être omise sous certaines conditions. En outre, le dzuun possède un certain nombre de MNFPs qui sont sémantiquement plus étroites que la MNFP par défaut *wāā*, telles que *dɛ̄* 'plus' et *kūrāā* 'jamais ; (pas) du tout'. Ces MNFPs spécifiques sont généralement utilisées seules, comme dans (9), remplaçant *wāā* tout comme *jamais* ou *nulle part* remplacent *pas* dans les exemples français (6) et (7). Cependant, à l'occasion, elles peuvent aussi être suivies de *wāā*, comme dans (10), ou elles peuvent coexister l'une avec l'autre lorsque le sens négatif doit être précisé davantage, comme dans (11). Enfin, certaines des formes qui fonctionnent comme des MNFPs spécifiques peuvent également apparaître dans des constructions positives, comme l'illustre *dɛ̄* en (12), où elle fonctionne comme une marque d'emphase. À cet égard, considérons le français *jamais*, qui peut également être utilisé dans des constructions positives, comme *si jamais* et *pour jamais*.

dzuun [1241]

(8) *à* ***náà*** *wù è tsí* ***wāā***
3SG NEG.PST bien 3SG.SBJV sauver NEG
'Il n'était pas bon qu'il soit sauvé.' (Solomiac 2007:270)

(9) *wó dɔ̀n náà, wó* ***nā*** *bómà jàá* ***dɛ̄***
2SG entrer venir.IPFV 2SG NEG sortie voir.IPFV plus
'Tu entres mais tu ne retrouves plus la sortie.' (Solomiac 2007:254)

(10) *tà bwèy, bɔ́ɔ́ rèè náà n'á rē yè ē séré kúráá wāā*
DEM moment vieux PL **NEG.PST** COP-3SG de 3PL.SBJV REFL prier **jamais** **NEG**
'En ce temps-là, les vieux ne voulaient jamais prier.' (Solomiac 2007:256, 578)

(11) *à* ***náà*** *fyā fyɛ̄* ***dɛ̄*** ***kūrāā***
3SG NEG.PST tissu blanc plus jamais
'[Quand la poule voulut venir avec le tissu blanc,] ce n'était plus du tout un tissu blanc.' (Solomiac 2007:539)

(12) *à cī, á! cī mún dzūnwēīnsíá mún sàn fìrìū* ***dɛ̄***
3SG QUO ah! QUO 1SG ami.DEF 1SG pied tromper.PFV EMPH
'Il dit : Ah! mon ami m'a bien trompé.' (Solomiac 2007:483)

Comme l'illustre l'exemple du dzuun, une marque ne doit pas nécessairement être une marque de négation dédiée (être intrinsèquement négative dans sa signification) pour être considérée comme une MNFP.

4.2.4 Ce que ça veut dire exactement d'être en fin de phrase

La description « en fin de phrase » dans MNFP fait référence à la position canonique de la marque de négation à la périphérie extrême droite d'une phrase. Une marque de négation donnée n'a pas besoin d'être dans la position finale absolue de la phrase dans toutes les constructions possibles pour être considérée comme une MNFP. Ce qui est pertinent, c'est que, dans la clause où le prédicat verbal est accompagné de deux ou plusieurs arguments nominaux simples et d'un complément simple modifiant le prédicat, tel qu'un adverbial simple de lieu ou de temps, la position de la marque de négation est déterminée par rapport à la phrase dans son ensemble et non par rapport au prédicat verbal, à ses arguments nominaux ou à son modifiant. Dans une langue donnée, la position de la MNFP par rapport aux autres marques de la périphérie droite et aux modifiants du prédicat verbal peut être fixe ou dépendre d'une série de facteurs, tels que

leur portée, leur signification, leur structure morphosyntaxique et leur longueur. Encore une fois, comme dans la discussion sur l'optionalité des MNFPs, une distinction claire et nette le long de ces lignes n'est pas toujours possible parce que le changement est graduel (bien que, plus souvent, la difficulté soit causée par le manque d'exemples pertinents dans les sources).

Un bon exemple des complexités possibles de la syntaxe des MNFPs est fourni par trois langues mande orientales du groupe boko-busa, le boko [boko1266], le busa [busa1253] et le bokobaru [boko1267], dont les MNFPs ont la forme *=o* (boko) et *=ro* (busa et bokobaru). Comme toutes les langues mande, les langues du groupe boko-busa ont un ordre strict des constituants SOVX dans les constructions transitives[12] et SVX dans les constructions intransitives, où X signifie « oblique », c'est-à-dire tout constituant (un argument ou un complément) autre que S et O (cf. Creissels 2005). La position canonique de la marque de négation *=(r)o* est en fin de phrase, comme l'illustre (13). Cependant, d'autres éléments de la périphérie droite ayant une portée phrastique, tels que la marque de question polaire *=à*, suivent la MNFP *=(r)o*, comme illustré dans (14). De plus, « sentence level adverbial phrases and clauses may follow the negative marker » (Jones 1998:299), comme illustré dans (15), qui peut être comparé à (14). La tendance des adverbiaux à suivre la marque de négation *=(r)o* est plus générale en busa et bokobaru, tandis qu'en boko, ce sont surtout les adverbiaux plus longs qui sont concernés. Enfin, la marque de négation *=(r)o* peut être suivie par le deuxième coordonné dans la construction de coordination alternative, comme dans (16). Ceci peut être analysé comme le résultat d'une ellipse, comme le fait Jones (1998:298). Alternativement, on peut y voir l'extraposition d'un constituant à la périphérie droite parce que les constituants lourds, comme ceux qui impliquent la coordination, sont évités dans les positions d'arguments (sujet, objet, phrase postpositionnelle). Cela ne serait pas du tout exceptionnel dans les langues mande et cela correspondrait également à la tendance à placer des adverbiaux plus longs après la MNFP *=(r)o*.

[12] Contrairement à la plupart des autres langues mande, les langues du groupe boko-busa autorisent également les objets nuls à interprétation anaphorique, mais seulement lorsque le référent est non humain et uniquement dans les constructions non perfectives, ainsi que dans les constructions perfectives avec des sujets nominaux ou un sujet pronominal de la troisième personne du pluriel (Jones 1998:212–213).

boko [boko1266]

(13) *‘í* ***ī*** *gbɛ́́* *pī-ɔ* *kã* *lá* *álɛ́* *‘e* *wà=**o***
liquide NEG.PFV personne ce-PL enivrer comme 2PL.PROG voir comme=NEG
‘La boisson n’a pas enivré ces gens comme vous le pensez.’ (Jones 1998:301)

(14) *‘àsí* *álɛ́* *ma* *nááí* *kɛ* *‘e* *tìa=**o**=à?*
alors 2PL.PROG 1SG.POSS confiance faire jusqu’à maintenant=NEG=PQ
‘Alors tu ne me fais toujours pas confiance?’ (litt. ‘Alors tu ne m’as pas fait confiance jusqu’à maintenant.’) (Jones 1998:299)

(15) *aa* *mɛ̀* *wa* *‘í* *mi=**o*** *‘e* *gɔɔ* *pɔ́* *wà*
3PL.PFV dire.PFV 3PL.LOG.FUT eau boire=NEG jusqu’à temps REL 3.INDF.PFV
aà *dɛ̀*
3SG.OBJ tuer.PFV
‘Ils ont dit qu’ils ne boiraient pas jusqu’au moment où il serait tué.’ (Jones 1998:299)

(16) *má* *‘ésé* *vĩ=**o*** *ge* *màsé*
1SG.STAT sorgho avoir=NEG ou maïs
‘Je n’ai pas de sorgho ni de maïs.’ (Jones 1998:299)

La position canonique de la marque de négation *=(r)o* dans le boko-busa est en fin de phrase et c’est ainsi qu’elle est classée dans cette typologie. En même temps, la variation synchronique observée dans son placement indique un processus diachronique en cours dans lequel la marque de négation est attirée vers la position immédiatement post-verbale.[13]

[13] L’attraction de la marque de négation en boko-busa, de sa position originale en fin de phrase vers la position immédiatement après le verbe, a probablement été déclenchée par l’influence du substrat du baatonum, une langue gur parlée immédiatement au sud-ouest du boko-busa. En baatonum, les marques de négation sont principalement préverbales, sauf dans la construction perfective négative, où la marque de négation préverbale est complétée par un suffixe verbal (Winkelmann & Miehe 2009:181–182). Dans les langues mande, la position des marques de négation varie, mais les MNFPs ne sont jamais attirées vers la position immédiatement après le verbe, car ce n’est pas la position associée au marquage de la polarité dans les langues mandé. En outre, il existe des raisons suffisantes pour supposer qu’une partie substantielle des populations boko-busa actuelles a passé du baatonum au boko-busa à un moment donné dans le passé. Par exemple, Jones (1998:5) souligne la relation évidente entre les termes boko-busa pour la partie de la population boko-busa non royale (‘paysans’, ‘vassaux’, ‘esclaves’) et les termes boko-busa pour les Baatonum.

4.2.5 Les MNFPs et l'ordre relatif de l'objet et du verbe

Contrairement à la typologie des marques de négation post-verbales de Dryer (2009), qui est limitée aux langues avec un ordre VO, l'ordre relatif de l'objet et du verbe n'est pas pertinent dans la typologie des MNFPs présentée ici. L'objet peut soit précéder le verbe comme dans les exemples dzuun et boko-busa ci-dessus, soit le suivre, comme dans l'exemple gbaya kara [gbay1283] (gbaya-manza-ngbaka) dans (17).

gbaya kara [gbay1283]

(17) *ʔám gbé sàdî hã́ kóò kóm ɲóŋ* ***ná***
1SG tuer\IPFV animal pour.que épouse POSS.1SG manger\IPFV NEG
'Je n'ai pas tué de gibier pour nourrir ma femme.' (litt. 'pour que ma femme mange')

Ce qui est pertinent pour ma typologie est que la position de la marque de négation sur la périphérie droite est déterminée par rapport à la phrase dans son entièreté. Les constructions avec ordre VO et les constructions avec ordre OV peuvent présenter différents types de problèmes analytiques pour déterminer si la marque de négation est en fin de phrase dans ce sens ou non.

Comme le souligne Dryer (2009:319), dans les langues africaines (subsahariennes) avec un ordre VO où la marque de négation suit l'objet, elle suit généralement aussi « any adverbs or adjunct phrases ». En d'autres termes, il s'agit typiquement d'une MNFP. A cet égard, les langues d'Afrique subsaharienne diffèrent des langues avec l'ordre VO et la marque de négation qui suit l'objet ailleurs dans le monde, comme l'allemand, la langue citée par Dryer (2009) comme exemple. Un exemple rare de langue d'Afrique subsaharienne similaire à l'allemand est le jur modo [jurm1239] (bongo-bagirmi; Andersen 1981; Persson & Persson 1991), parlé au Soudan du Sud, à la périphérie de l'aire principale du trait MNFP. Le jur modo utilise l'ordre SVX dans les constructions intransitives et l'ordre SVOX dans les constructions transitives. La position immédiatement à la fin de la phrase verbale, c'est-à-dire, après V dans la construction intransitive et après O dans la construction transitive ou, formulé différemment, immédiatement avant la position X, semble être réservé en jur modo pour au moins deux marques grammaticales, dont l'une est la marque de négation *dé*, comme illustré dans (18) et (19), et l'autre est la marque du résultatif ou parfait *dɛ̃ní* (appelé

« perfective » par Andersen 1981 ou « completive » par Persson & Persson 1991), comme illustré dans (20).[14]

jur modo [jurm1239]

(18) *m-údɔ̂ ndɔ̀bɔ̀ **dé** kpɛ̀ tí=ì*
1SG-faire travail NEG de.nouveau avec=2SG
'Je ne travaillerai plus avec toi.' (Andersen 1981:59)

(19) *mɔ̀rɔ́ ílábá **dé** rɔ̀ kòbì*
lance tomber NEG à buffle
'La lance n'a pas touché le buffle.' (Andersen 1981:80)

(20) *kìrábà ùpɛ̀ kúmú **dɛ́ní** dì mì màlìbìwù*
chacal libérer lièvre PRF de dans piège
'Le Chacal a libéré le Lièvre du piège.' (Persson & Persson 1991:15)

Outre la situation typique dans les langues VO africaines, où le statut en fin de phrase d'une marque de négation est relativement évident, nous trouvons également un certain nombre de langues VO à la périphérie de l'aire principale du trait MNFP, où une marque de négation gravite vers la fin de la phrase mais où il n'est pas évident si sa position canonique doive être caractérisée comme en fin de phrase ou non. L'un des exemples les plus clairs d'une telle langue est le nzadi [nzad1234] (bantu B865; Crane, Hyman & Tukumu 2011), dont la description fournit un aperçu détaillé de la syntaxe de la marque de négation post-verbale. En nzadi, la négation est marquée en deux positions dans la phrase : la première marque se trouve avant le verbe « in the auxiliary » (la forme de cette marque de négation dépend des valeurs du TAM) et la seconde marque, *bɔ*, se trouve après le verbe « towards the end of the clause » et prend une portée « over any of the elements » de la phrase (Crane, Hyman & Tukumu 2011:169, 173). (21) schématise les positions possibles de *bɔ* dans diverses structures phrastique.

[14] Contrairement à Dryer (2009:320) qui affirme que la marque de négation « can be freely positioned among adverbial or adjunct elements », Andersen (1981) et Persson & Persson (1991) précisent seulement que la marque de négation est un type d'adverbe qui doit se trouver « in the adjunct [position], separated from the verb by the object » (Persson & Persson 1991:15). Pourtant, dans tous les exemples trouvés dans ces sources, le marqueur de négation est toujours le premier des adverbes ou circonstants, immédiatement après le verbe ou, s'il est présent, l'objet.

nzadi [nzad1234]

(21) Les positions possibles de la marque de négation post-verbale *bɔ* (Crane, Hyman & Tukumu 2011:171)[15]

S-V-*bɔ*	*S-*bɔ*-V	**bɔ*-S-V
S-V-O-*bɔ*	?S-V-*bɔ*-O	
S-V-IO-DO-*bɔ*	S-V-IO-*bɔ*-DO	*?S-V-*bɔ*-IO-DO
S-V-DO-Obl-*bɔ*	?S-V-DO-*bɔ*-Obl	*?S-V-*bɔ*-DO-Obl
S-V-DO-Obl_{BEN}-*bɔ*	S-V-DO-*bɔ*-Obl_{BEN}	*S-V-*bɔ*-DO-Obl_{BEN}
S-V-X-*bɔ*	S-V-*bɔ*-X	

Pour les besoins de ma typologie, les marques de négation similaires à nzadi *bɔ* sont classées comme optionnellement en fin de phrase. D'un point de vue diachronique, une telle indétermination suggère un changement syntaxique en cours par lequel une marque de négation qui, en vertu de son étymologie, a évolué à l'origine dans une certaine position dans la structure de la phrase, est attirée vers une position différente dans la

[15] L'astérisque <*> marque les options non grammaticales. Ailleurs dans la source, les exemples d'options de placement de *bɔ* marqués par la combinaison <*?> sont également caractérisés comme non grammaticales, il reste donc peu clair quelle différence entre <*> et <*?> était entendue par les auteurs dans ce tableau. Le point d'interrogation <?> marque les options qui sont caractérisées comme « strongly dispreferred » ou « at least marginally acceptable ». *S-V-IO-DO* représente les constructions ditransitives à double objet, où l'objet indirect n'est pas marqué. *S-V-DO-Obl* désigne les constructions ditransitives à objet indirect, où l'objet indirect, dit oblique, est introduit par la préposition locative *kó*. J'ai ajouté au tableau original la ligne avec l'oblique bénéfactif (Obl_{BEN}) marqué par *sám ˈé N* (litt. 'la raison de N'), car il diffère des obliques introduits par la préposition *kó* en ce que « the preferred ordering may place *bɔ* before the benefactive », bien que sans aucune « strong preference either way » (Crane, Hyman & Tukumu 2011:170). Enfin, « *X* can be a non-object complement, or any adjunct, and may co-occur with direct and indirect objects [including obliques], with *bɔ* placement restricted with regard to objects as in other cases » (Crane, Hyman & Tukumu 2011:171).

structure de la phrase, probablement parce que cette position-ci est associée à l'expression de certains types de sens.[16]

Les langues africaines avec un ordre OV et une marque de négation post-verbale peuvent être subdivisées en deux groupes pour les besoins de ma typologie. Dans le premier groupe, le verbe est normalement suivi par des constituants (arguments ou circonstants) autres que l'objet. La plupart de ces langues semblent se comporter comme les langues mande dzuun et boko présentées dans §4.2.3 et §4.2.4, dans le sens que la marque de négation post-verbale suit également d'autres constituants post-verbaux et peut donc être qualifiée de marque en fin de phrase. La plupart de ces langues sont en fait les langues mande. Dans le deuxième groupe, la phrase se termine essentiellement par le verbe, de sorte que la question si la marque de négation post-verbale est positionnée par rapport à la phrase dans son entièreté ou seulement par rapport au verbe n'a pas vraiment beaucoup de sens. Cependant, lorsque des preuves diachroniques sont disponibles, il est généralement clair que la marque de négation post-verbale est positionné par rapport au verbe et non par rapport à la phrase, car elle provient souvent d'un verbe principal réanalysé comme auxiliaire (cf. van Gelderen 2008:232–233; Lucas 2009 sur les langues afro-asiatiques). Par conséquent, par défaut, les marques de négation post-verbales de ces langues ne sont pas caractérisées comme étant en fin de phrase aux fins de ma typologie. Cette situation est courante dans les langues afro-asiatiques d'Afrique du nord et de l'est (cushitique, omotique, sémitique), comme l'illustre (22) du daasanach [daas1238] (cushitique ; Tosco 2001), dans certains groupes nilo-sahariens du Tchad et du Soudan (comme saharien, four et nubien) et dans les langues dogon et ijoïde d'Afrique de l'ouest, comme l'illustre (23) de jamsay [jams1239] (dogon ; Heath 2008).

[16] La position originale de la marque *bɔ* du nzadi est probablement après l'objet indirect, soit celui qui n'est pas marqué, soit celui introduit par la préposition *kó* ou, en l'absence d'un tel objet indirect, après l'objet direct ou, quand aucun objet n'est présent, après le verbe. C'est-à-dire qu'elle est maintenant attirée vers la position en fin de phrase, sans doute en raison de sa portée phrastique par défaut et de sa fonction intermédiaire d'atténuateur de la force assertive de la prédication négative dans son entièreté. Sa position originale peut être expliquée par son étymologie probable en tant que pronom possessif, qui était coréférentiel avec le sujet et fonctionnait comme une sorte d'atténuateur, quelque chose comme 'quant à X_i [S_i ne P pas]', qui peut être grossièrement comparé à certains usages des pronoms emphatiques en français, comme dans *Pierre ne sait pas, lui* (alors que d'autres pourraient savoir)'. Compte tenu de la forme de *bɔ*, c'est très probablement la forme de la troisième personne du pluriel qui s'est généralisée. Comme décrit par Devos & van der Auwera (2013), les pronoms possessifs ne sont pas rares comme source de marques de négation secondaires dans les langues bantu de la région.

daasanach [daas1238]

(22) *yáa ʔúm **ma** ká šuggun-**iŋ***
1SG.SBJ enfants NEG ici amener.IPFV-NEG
'Je ne vais pas amener les enfants ici.' (Tosco 2001:299)

jamsay [jams1239]

(23) *ɔ́ɣɔ́rɔ́ kò-rú yɔ̀wɔ̀-**l**-á*
rapidement NONHUM-DAT accepter-PFV.NEG-3PL.SBJ
'Ils ne l'ont pas accepté facilement [= charrue].' (Heath 2008:368)

4.3 Les données

Les données de cette étude sur les MNFPs proviennent de descriptions grammaticales individuelles complétées par un certain nombre d'enquêtes typologiques existantes sur le marquage de négation en Afrique, telles que l'enquête de Dryer (2009) sur les marques de négation post-verbales dans les langues VO d'Afrique centrale, l'étude de Devos & van der Auwera (2013) sur les marques de négation multiples dans les langues bantu[17] et l'étude de Beyer (2009) sur les marques de négation doubles dans les langues de la région centrée sur le bassin de la Volta en Afrique occidentale. J'ai essayé de recouper les informations provenant des enquêtes typologiques dans les descriptions grammaticales chaque fois que cela était possible.

Mon échantillon se compose de 618 langues, dont 256 utilisent une sorte de MNFP, tandis que 328 langues n'en ont clairement pas et que, pour 34 langues, les informations disponibles n'étaient pas suffisantes pour prendre une décision éclairée. Dans la plupart des cas, j'ai combiné ces deux derniers groupes en tant que langues sans MNFPs (362 langues). La distribution géographique des 618 langues de mon échantillon a été présentée dans la figure 13 dans §4.1 que je reproduis ici comme figure 14. La figure 14 représente également cette distribution sous forme d'intensité spatiale, c'est-à-dire le degré de concentration des langues prises comme points dans l'espace. La plus importante concentration de langues se trouve dans la zone autour de la frontière entre le Cameroun et le Nigeria. Une autre zone de forte concentration de langues s'étend du Togo au sud-ouest du Burkina Faso. La figure 15 montre la distribution géographique et l'intensité spatiale des 256 langues qui ont des MNFPs. Le profil général de distribution des langues avec MNFPs de la figure 15 ressemble au profil de distribution de l'échantillon dans son ensemble de la figure 14. Le profil de la figure 15 est cependant

[17] Je suis reconnaissant à Maud Devos et Johan van der Auwera de m'avoir donné accès à la base de données source qu'ils ont créée pour cette enquête.

plus circonscrit dans l'espace, dans presque toutes les directions. Il est essentiellement limité au nord de l'Afrique subsaharienne. Sa zone focale, bien qu'également située dans la région autour de la frontière entre le Cameroun et le Nigeria, a une position relativement septentrionale et son extension vers l'ouest, vers le sud-ouest du Burkina Faso, est un peu moins prononcée et a une orientation latitudinale plus clairement est-ouest (par opposition à une orientation plus sud-est-nord-ouest dans la figure 14).

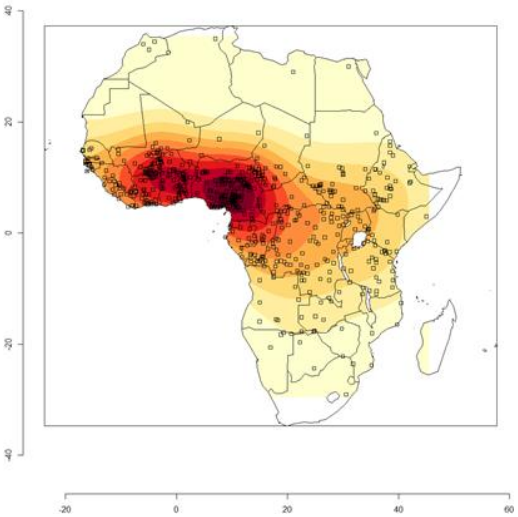


FIGURE 14. La distribution géographique des 618 langues de l'échantillon et un graphique de leur intensité spatiale.

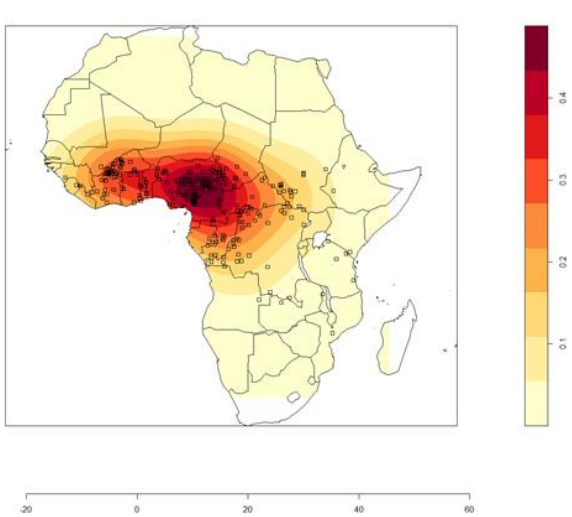


FIGURE 15. La distribution géographique des 256 langues avec MNFPs et un graphique de leur intensité spatiale.

La figure 16 montre la distribution géographique et l'intensité spatiale des 362 langues qui n'ont pas de MNFPs.

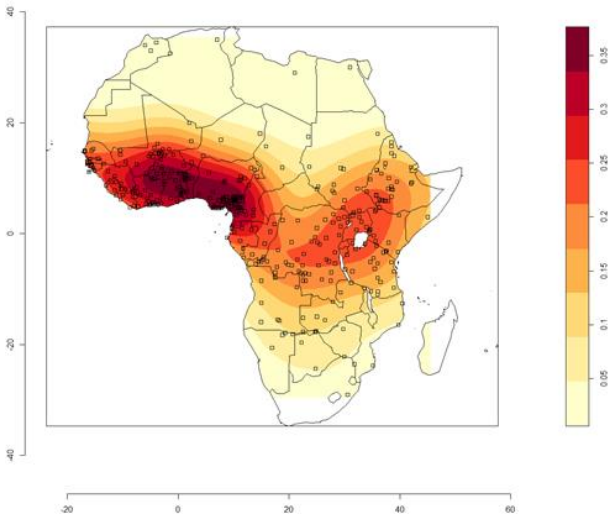


FIGURE 16. La distribution géographique des 362 langues sans MNFPs et un graphique de leur intensité spatiale.

Le schéma de distribution des langues sans MNFPs de la figure 16 est assez différent, surtout dans sa partie orientale. Tout d'abord, la distribution de la figure 16 est beaucoup plus étendue. Il est important de noter qu'elle est caractérisée par une nette dépression en Afrique centrale, où le profil présente une courbe en forme de U. Cette dépression dans le profil de la figure 16 est créée par la présence d'un groupe de langues relativement homogène avec des MNFPs dans cette région, dont l'importance n'est peut-être pas évidente à partir de la figure 15. La dépression (et le groupe de langues avec MNFPs qui en est responsable) est orientée nord-est-sud-ouest et s'étend de la République centrafricaine le long du fleuve Congo, correspondant sur la carte à la frontière entre le Congo et la République démocratique du Congo. Cette dépression fait que la forme de la distribution descend vers le sud au Cameroun, parallèlement à la côte, vers le cours inférieur du fleuve CongErreo, où elle tourne ensuite vers l'est et remonte finalement autour du centre de la République démocratique du Congo en direction du nord-est, vers l'Éthiopie. Il est intéressant de comparer la zone focale ouest-africaine de la figure 16 avec celle de la figure 15. La zone focale de la figure 16 est à la fois plus large et plus prononcée, s'étendant dans une orientation sud-est-nord-ouest similaire à celle que nous trouvons dans l'ensemble de l'échantillon dans la figure 14. Comme pour la zone focale de la figure 15, l'extrémité orientale de la zone focale de la figure 16 est située autour de la frontière entre le Cameroun et le Nigeria, mais sa position est nettement plus méridionale, s'étendant du sud-est du Nigeria au sud du Cameroun.

4.4 Au-delà du binaire : augmenter la granularité des données

Les profils de répartition géographique des langues de l'échantillon présentés dans les figures 14, 15, et 16 dans §4.3 sont essentiellement des motifs de points. En tant que tels, ils nous montrent l'étendue globale des langues avec et sans MNFPs et mettent en évidence les régions de forte et de faible concentration des deux types de langues. Il s'agit d'informations précieuses pour une première approche du phénomène. Cependant, cette représentation binaire cache une diversité importante parmi les langues avec MNFPs. Pour mieux saisir cette diversité et ainsi avoir une meilleure idée des dynamiques historiques et spatiales possibles qui ont conduit à la situation actuelle, nous devons augmenter le degré de granularité de nos données. Suite à la discussion des différentes questions liées à la délimitation du trait MNFP dans §4.2, je vais utiliser deux paramètres. Le premier est le caractère obligatoire des MNFPs, c'est-à-dire si les MNFPs sont obligatoires ou optionnelles, et le second, que j'appellerai *la liberté par construction*, tient compte de la possibilité d'existence des restrictions sur l'utilisation des MNFPs dans différentes constructions. Le tableau 1 résume les combinaisons

possibles des valeurs de deux paramètres et un schéma de codage avec les valeurs pseudo-numériques qui leur sont attribuées.[18] La dernière colonne indique le nombre de langues de chaque type. Dans le schéma de codage utilisé, les MNFPs qui sont obligatoires et qui ne sont pas soumises aux restrictions sur l'utilisation en fonction de construction sont classés au plus haut niveau, soit 4, tandis que les MNFPs qui sont à la fois soumises aux restrictions sur l'utilisation en fonction de construction et sont optionnelles sont classées au plus bas niveau, soit 1.[19]

Caractère obligatoire	Liberté par construction	Schéma de codage	Nombre de langues
pas de MNFPs		0	328
pas clair		0.5	34
optionnel	avec restrictions	1	7
optionnel	sans restrictions	2	22
obligatoire	avec restrictions	3	31
obligatoire	sans restrictions	4	196

TABLEAU 1. Les combinaisons possibles des valeurs de deux paramètres, à savoir [le caractère obligatoire] et [la liberté par construction], et le schéma de codage avec les valeurs pseudo-numériques qui leur sont attribuées.

En principe, l'un ou l'autre des deux paramètres pourrait être classé en premier et il se trouve que, pour cette distribution particulière de langues avec des MNFPs, qui est fortement déséquilibrée d'un côté, les deux options produisent des résultats très similaires. Néanmoins, j'ai une préférence de principe pour le classement du paramètre de caractère obligatoire au premier rang, car je conçois le caractère obligatoire comme la propriété déterminante des significations grammaticales (voir Idiatov 2008 pour une

[18] Les valeurs sont pseudo-numériques dans le sens où leurs valeurs numériques sont fondamentalement arbitraires et sont juste destinées à refléter l'ordre relatif des différentes combinaisons des paramètres. En fait, que nous utilisions ces valeurs numériques ou simplement une liste ordonnée de facteurs n'a pas beaucoup d'importance. Les deux méthodes donnent des résultats très similaires en termes d'analyse spatiale, par exemple lorsque nous les visualisons en utilisant l'interpolation spatiale. Cependant, je préfère utiliser les valeurs pseudo-numériques car elles permettent de mieux saisir le statut relatif des langues qui n'ont pas de MNFPs par opposition aux différents types de langues qui ont ou pourraient avoir des MNFPs.

[19] J'ai également essayé un schéma de codage alternatif avec les valeurs pseudo-numériques compressées sur une échelle de 0 à 1, la catégorie « pas clair » positionnée au milieu avec la valeur 0,5, et les autres catégories ayant les valeurs 0,625, 0,75 et 0,875 respectivement. L'espoir était que ce schéma alternatif serait mieux à même de rendre compte de la hiérarchie entre les six valeurs du trait MNFP que le premier schéma. Toutefois, il s'est avéré que le choix entre les deux schémas n'a pas d'effet notable sur les visualisations.

discussion détaillée). L'absence de restrictions quant à l'utilisation d'un marqueur grammatical en fonction de construction est une propriété des marqueurs grammaticaux canoniques (au sens de la typologie canonique ; voir Brown, Chumakina & Corbett 2013). Par conséquent, les MNFPs qui sont à la fois obligatoires et exempts de restrictions en fonction de construction sont des marqueurs grammaticaux canoniques, tandis que d'autres types de MNFPs n'atteignent pas ce statut.

Un point important à mentionner en ce qui concerne la classification du tableau 1 est qu'elle classe les langues, et non les MNFPs. Si une langue possède plusieurs MNFPs qui diffèrent par rapport aux deux paramètres, je choisis la MNFP la plus hautement cotée, comme étant la plus proche du statut de marqueur grammatical canonique, pour représenter la langue dans son entièreté. Certes, de cette façon, une partie de la diversité n'est pas correctement reflétée dans la typologie, mais je ne vois pas très bien comment je peux intégrer cette information. De plus, j'ai aussi l'impression que l'ajouter n'aurait pas d'effets significatifs sur les résultats globaux.

4.5 Une typologie aréale des MNFPs en Afrique subsaharienne

Dans cette section, je discute d'abord les résultats et les pièges potentiels de deux méthodes d'analyse spatiale et de visualisation de la distribution des valeurs du trait MNFP en Afrique subsaharienne, ainsi que de certaines corrélations géographiques qui ressortent de cette analyse (§4.5.1). En particulier, j'applique l'interpolation spatiale (§4.5.1.1) en utilisant deux types de lissage différents, le lissage à noyau et le lissage pondéré par la distance inverse, et la modélisation additive généralisée (§4.5.1.2). Les différentes méthodes utilisées convergent vers le même profil spatial du trait MNFP. Elles confirment l'existence, la position et la forme globale de deux aires de convergence, l'Aire de Convergence Centrale couvrant l'est de l'Afrique de l'Ouest et certaines parties de l'Afrique Centrale et l'Aire de Convergence Occidentale limitée à l'Afrique de l'Ouest. Les deux aires sont séparées par une discontinuité majeure autour du Ghana, du Togo et du Bénin. Des deux aires de convergence, l'Aire de Convergence Centrale peut être qualifiée d'aire de convergence primaire, étant donné sa proéminence, et l'Aire de Convergence Occidentale d'aire de convergence secondaire. Dans §4.5.2, j'aborde la question du foyer historique de l'Aire de Convergence Centrale. Je soutiens en particulier que, malgré la proéminence apparente d'une zone située dans le sud du Tchad et de la République centrafricaine au sein de l'Aire de Convergence Centrale, elle ne peut pas représenter son foyer historique et qu'il est beaucoup plus probable que le foyer historique primaire de l'Aire de Convergence Centrale soit situé immédiatement au nord-ouest de la République centrafricaine, le long du corridor du fleuve Benue qui

va du sud du Tchad au centre du Nigeria en passant par le nord du Cameroun. En même temps, comme je le discute dans §4.5.3, cette zone du sud du Tchad et de la République centrafricaine, qui occupe une place importante au sein de l'Aire de Convergence Centrale, a dû servir de source pour la diffusion du trait MNFP parmi les langues bantu plus au sud, dans le couloir du fleuve Congo et au nord de la République démocratique du Congo. La section §4.5.3 propose en outre une discussion sur la question plus large de la distribution des MNFPs optionnelles et/ou soumises aux restrictions sur l'utilisation en Afrique et soutient que, comme prévu, ces MNFPs grammaticalement non canoniques ont tendance à être également périphériques du point de vue aréale.

4.5.1 L'analyse spatiale

La distribution spatiale des langues avec différents types de MNFPs (tels que distingués dans §4.4) et des langues sans MNFPs peut être inspectée de plusieurs façons. Ainsi, je les visualise d'abord à l'aide de l'interpolation spatiale (§4.5.1.1) et après à l'aide de la modélisation additive généralisée (§4.5.1.2). Dans les deux cas, je le fais en utilisant les valeurs pseudo-numériques décrites dans §4.4.

4.5.1.1 L’interpolation spatiale

La figure 17 montre le résultat de l’interpolation spatiale utilisant le lissage par noyau et la figure 18 montre le résultat de l’interpolation spatiale utilisant le lissage par pondération inverse à la distance.

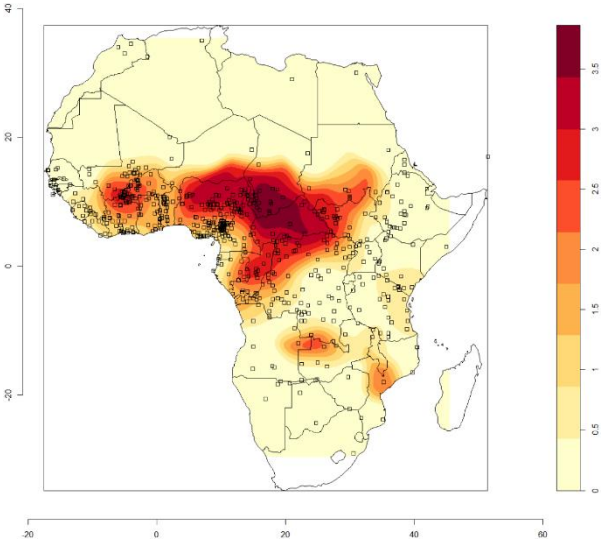


FIGURE 17. La visualisation par l’interpolation spatiale du trait MNFP (comme décrit dans §4.4) dans un échantillon de 618 langues à l’aide du lissage par noyau Gaussien (la valeur par défaut de la bande passante ajustée par 1,3).

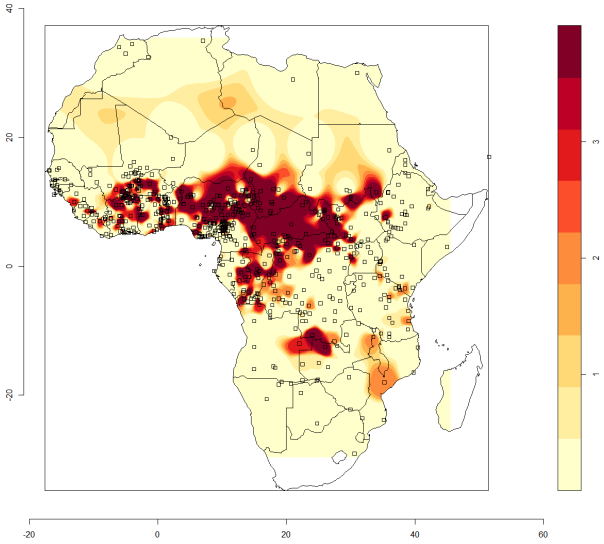


FIGURE 18. La visualisation par l’interpolation spatiale du trait MNFP (comme décrit dans §4.4) dans un échantillon de 618 langues à l’aide de la pondération inverse à la distance (la puissance = 6).

Comme j’ai discuté dans §3.3.2.2, en fonction de la manière exacte dont les données sous-jacentes sont distribuées dans l’espace, l’interpolation spatiale peut produire certains artefacts de visualisation dont il faut être conscient lors de l’analyse des résultats et il est évident que ces deux figures n’en font pas exception. Pour une discussion détaillée des artefacts de visualisation qui affectent précisément ces deux graphiques d’interpolation spatiale je renvoie à §3.3.2.2.

En outre, comme j’ai également discuté dans §3.3.2.2, les deux méthodes d’interpolation spatiale produisent des visualisations légèrement différentes, ce qui peut permettre de mieux mettre en évidence différents aspects de la distribution spatiale. Ainsi, l’interpolation utilisant le lissage par noyau dans la figure 17 permet de visualiser mieux la structure globale de la distribution spatiale du trait MNFP. Elle montre clairement une discontinuité majeure dans la distribution des langues avec des MNFPs en Afrique subsaharienne septentrionale autour du Ghana, du Togo et du Bénin. Cette discontinuité sépare une aire de convergence du trait MNFP secondaire qui est centrée sur la région où les frontières du Burkina Faso, du Mali et de la Côte d’Ivoire se rejoignent. Par souci de commodité de référence, j’appelle cette aire de convergence du

trait MNFP secondaire en Afrique subsaharienne septentrionale l'Aire de Convergence Occidentale et l'aire de convergence du trait MNFP principale à l'est de celle-ci l'Aire de Convergence Centrale. L'interpolation utilisant le lissage par noyau de la figure 17 et l'interpolation utilisant le lissage par pondération inverse à la distance de la figure 18 montrent qu'en Afrique subsaharienne septentrionale les régions caractérisées par la présence du trait MNFP sont largement confinées à l'arrière-pays, donc pas sur les côtes. La localisation des trois extensions les plus nettes la zone de concentration du trait MNFP vers la côte du golfe de Guinée est un peu mieux visible sur la figure 18. Ainsi, la première de ces extensions se trouve autour du sud du Togo et du Bénin, franchissant largement la discontinuité entre l'Aire de Convergence Centrale et l'Aire de Convergence Occidentale. La deuxième extension côtière est située dans le sud du centre du Nigeria et est formée par la diffusion vers le sud des langues edoïdes. Si les deux premières extensions côtières sont elles-mêmes susceptibles de résulter d'événements relativement récents de diffusion des langues et/ou de contact linguistique, la discontinuité qui les sépare peut tout aussi bien être accidentelle. Ainsi, la zone de discontinuité est occupée par une grande langue (ou un groupe de langues étroitement liées), à savoir le yoruba, qui a dû s'étendre dans la zone de discontinuité à partir d'une région plus à l'intérieur du Nigeria central relativement récemment et il se peut que la proto-langue respective simplement n'avait pas le trait MNFP par hasard ou l'ait perdu après étant entrée dans la zone. À cet égard, il faut noter que l'on trouve des MNFPs dans diverses langues apparentées parlées juste à l'extérieur de la zone de discontinuité, comme de nombreuses langues edoïdes ou l'igala, qui appartient au même groupe linguistique de niveau inférieur, à savoir le groupe yoruboïde, que le yoruba. La troisième extension côtière le long du couloir du fleuve Congo est due à des mouvements linguistiques et/ou de populations relativement récents en provenance d'Afrique centrale, qui ont affecté les langues bantu de cette région (voir §4.5.3).

4.5.1.2 La modélisation additive généralisée

La figure 19 montre le résultat de la modélisation additive généralisée sous forme de tracé de contours avec le schéma de couleurs de la carte thermique. Une couleur plus claire correspond à une valeur pseudo-numérique plus élevée du trait MNFP comme décrit dans §4.4. Les lignes de contour sont des isoplèthes qui marquent les déviations de la moyenne en termes d'écart type.

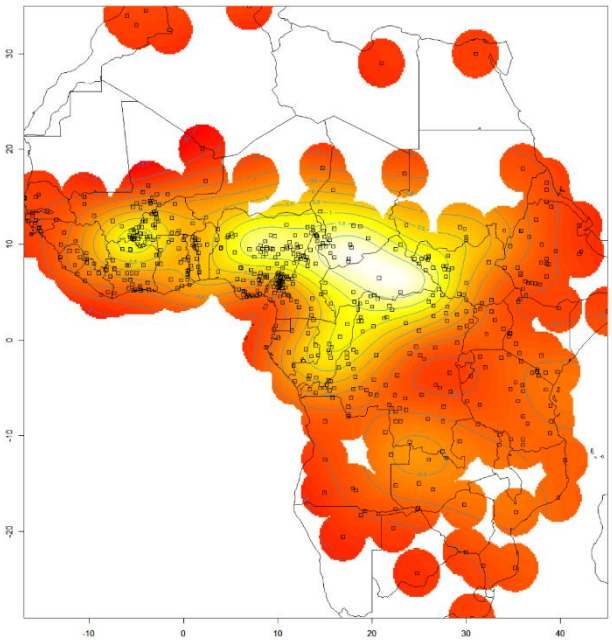


FIGURE 19. Le tracé de contours avec le schéma de couleurs de la carte thermique représentant la surface de régression GAM des valeurs pseudo-numériques du schéma de codage du trait MNFP décrit dans §4.4 en fonction de la combinaison de la longitude et de la latitude en utilisant des splines de régression à plaque mince. Le résumé du modèle : k = 13 (k-index = 0.95, p = 0.065, k′ = 168), fonction = gaussienne, edf = 41.73, la déviance expliquée = 43.7%, AIC = 2234, intercept = 1.53, p < 2e-16.

Le graphique GAM de la figure 19 est très similaire aux graphiques d'interpolation spatiale de la figure 17 et de la figure 18, moins les artefacts de visualisation accidentels produits par l'interpolation spatiale et discutés dans §3.3.2.2. Ainsi, les différentes méthodes utilisées convergent vers le même modèle spatial du trait MNFP. Elles confirment l'existence, la position et la forme globale de deux aires de convergence, l'Aire de Convergence Centrale et l'Aire de Convergence Occidentale, séparées par une discontinuité majeure au niveau du Ghana, du Togo et du Bénin.

Des deux aires de convergence, l'Aire de Convergence Centrale peut être qualifiée d'aire de convergence principale, étant donné sa proéminence, et l'Aire de Convergence Occidentale d'aire de convergence secondaire. La forme générale des deux aires de convergence peut être décrite comme l'arrière-pays du golfe de Guinée. Partant de l'Aire de Convergence Occidentale au sud du Mali et au nord de la Côte d'Ivoire, la région où le trait MNFP est bien présent s'étend vers l'est, suivant principalement les grasslands

et les savanes boisées au nord de la zone forestière, restant pour la plupart en dehors des régions côtières. Elle n'est interrompue que par une seule discontinuité majeure séparant l'Aire de Convergence Occidentale de l'Aire de Convergence Centrale. Géographiquement, cette discontinuité majeure correspond assez bien à ce que l'on appelle le Dahomey Gap, un corridor mélange de forêt sèche et de savane descendant vers le sud qui interrompt la forêt tropicale zonale ouest-africaine. Cela peut sembler étrange au premier abord, car l'Aire de Convergence Occidentale et l'Aire de Convergence Centrale se trouvent elles-mêmes dans la zone de savane. Cependant, l'orientation nord-sud de ce corridor de savane peut également être considérée comme propice à l'interruption de la dynamique générale est-ouest des mouvements des populations et des langues dans la partie occidentale de l'Afrique subsaharienne septentrionale. Ainsi, je pense que cette discontinuité est principalement due à l'effet combiné de la propagation vers le sud du songhay dans la zone actuelle de la brèche depuis le nord et de la propagation vers le nord du sous-groupe tano de la famille linguistique kwa depuis les régions côtières au sud.

La légère inclinaison vers le sud-est de l'Aire de Convergence Centrale dans le sud du Tchad en direction de la République centrafricaine suit bien l'orientation des zones écologiques, de la topographie et de l'hydrographie de cette partie de l'Afrique subsaharienne septentrionale, comme l'illustre la figure 20 sur une carte en relief de l'Afrique.[20] L'Aire de Convergence Centrale déborde en outre marginalement sur le Soudan du Sud à l'est et, de manière beaucoup plus significative, sur l'Afrique centrale équatoriale au sud-ouest, le long du corridor du fleuve Congo (voir §4.5.3). Les deux zones plausibles à travers lesquelles l'interaction entre l'Aire de Convergence Centrale et ces régions voisines est susceptible d'avoir eu lieu sont marquées dans la figure 20

[20] Les frontières orientales du Tchad et de la République centrafricaine correspondent en grande partie à la ligne de partage qui sépare les bassins versants du lac Tchad et du fleuve Congo, à l'ouest, du bassin versant du fleuve Nil, plus à l'est. La frontière sud-ouest du Tchad et la frontière ouest de la République centrafricaine reflètent approximativement la ligne de partage séparant les mêmes bassins versants du lac Tchad et du fleuve Congo des bassins versants du fleuve Niger et de quelques petits fleuves se jetant dans le golfe de Guinée. La forme des lignes de partage résulte naturellement du relief de la région.

comme A et B respectivement, la différence de degré d'interaction étant représentée graphiquement par la différence de style de police des deux symboles.

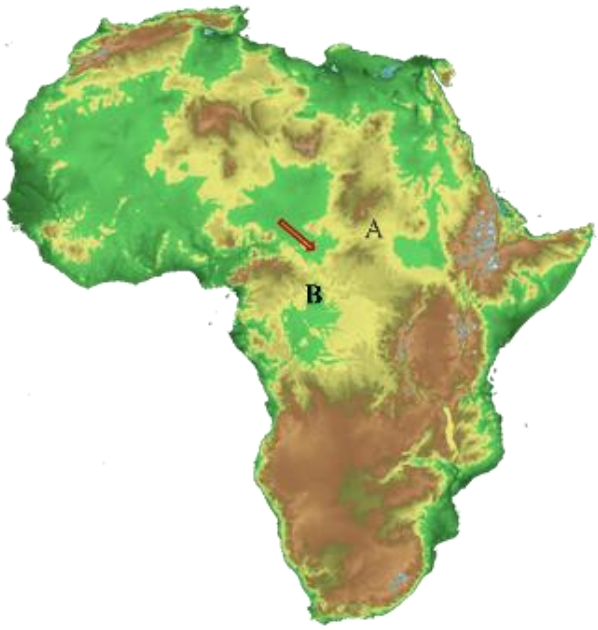


FIGURE 20. Une carte en relief de l'Afrique : la flèche met en évidence la corrélation entre la topographie et l'inclinaison vers le sud-est de l'Aire de Convergence Centrale autour du sud du Tchad. Les lettres A et B marquent les zones d'interaction primaire plausibles entre l'Aire de Convergence Centrale et les régions voisines du Soudan du Sud et de l'Afrique centrale équatoriale. La différence de style de police des deux symboles représente graphiquement la différence de degré d'interaction.

4.5.2 Le foyer historique de l'Aire de Convergence Centrale

Dans la figure 17, qui est un tracé d'interpolation spatiale avec un lissage par noyau, et dans la figure 19, qui est une visualisation d'un modèle additif généralisé, l'Aire de Convergence Centrale est caractérisée par une région très proéminente dans le sud du Tchad et en République centrafricaine (comme le reflète son ombrage plus foncé dans la figure 17 et sa couleur plus claire dans la figure 19). Il peut donc être tentant d'interpréter la région du sud du Tchad et de la République centrafricaine comme le noyau ou le foyer de l'Aire de Convergence Centrale, avec toutes les implications historiques évidentes qu'une telle interprétation impliquerait. En même temps, cette région très saillante est absente de la figure 18, qui utilise une méthode d'interpolation spatiale différente, à savoir le lissage par pondération inverse à la distance. De même, la même région est tout sauf proéminente dans la figure 15, qui montre la distribution des langues avec des MNFPs et leur intensité spatiale. Il est clair qu'il faut faire preuve de prudence lorsqu'on interprète la pertinence de cette région très proéminente au sein de l'Aire de Convergence Centrale. En fait, je pense que la proéminence de la région dans le sud du Tchad et en République centrafricaine est un épiphénomène et que cette région

n’est pas le foyer historique de l’Aire de Convergence Centrale. Un bien meilleur candidat pour ce rôle est le corridor du fleuve Benue qui va du centre du Nigeria au sud du Tchad en passant par le nord du Cameroun. L’importance apparente de cette région au sein de l’Aire de Convergence Centrale résulte de l’effet combiné d’un certain nombre de facteurs, qui concernent principalement sa partie sud, en République centrafricaine. D’une part, nous avons la géographie de la région, qui en fait une sorte de cul-de-sac avec beaucoup de zones marécageuses et saisonnièrement inondées. D’autre part, nous avons l’histoire linguistique et démographique de la région, qui semble être caractérisée par de forts effets fondateurs.

En examinant de plus près cette partie de l’Aire de Convergence Centrale, il est tout d’abord important de noter que cette région de l’Afrique centrale est très homogène en ce qui concerne le trait MNFP, la plupart des langues (sinon toutes) ayant des MNFPs canoniques (type 4). En même temps, cette région d’Afrique centrale est à la fois assez homogène sur le plan linguistique et assez peu peuplée. Elle est faiblement peuplée à la fois en termes absolus, comme le montre la faible densité de lieux habités dans cette région sur la figure 21, et en termes de langues qui y sont parlées. Ce dernier point est manifeste dans la figure 13, qui montre la répartition des langues de l’échantillon.

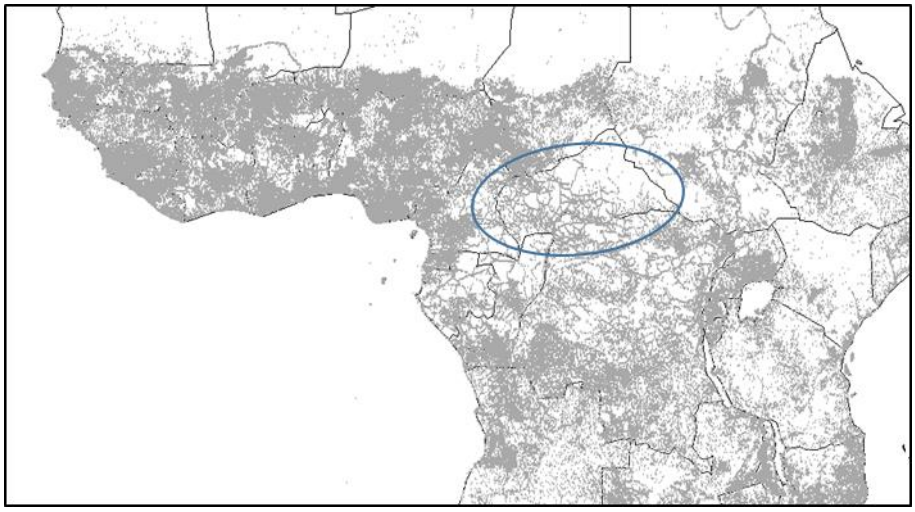


FIGURE 21. Lieux habités en Afrique subsaharienne septentrionale (sur la base des données de GeoNames.org) – l’ovale centré sur la République centrafricaine indique approximativement la partie méridionale de l’Aire de Convergence Centrale.

Pour voir comment la faible densité linguistique peut être pertinente pour la proéminence apparente de la région en question, rappelons la discussion de certains des artefacts possibles de la visualisation au moyen de l’interpolation spatiale dans §3.3.2.2. Quant à l’homogénéité linguistique, cette région est occupée par un petit nombre de groupes linguistiques, tous plutôt de faible profondeur, à savoir les langues gbaya-manza-ngbaka, les langues sere-ngbaka-mba, les langues banda, les langues ngbandi-mongoba-kazibati, les langues zande et les langues sara-bongo-bagirmi occidentales.

Tous ces groupes, à l'exception du dernier, ont traditionnellement été classés ensemble sous l'étiquette des langues oubangiennes, bien que plus récemment les langues gbaya-manza-ngbaka aient été exclues de ce groupe. Le groupe sara-bongo-bagirmi occidental est une branche de la famille sara-bongo-bagirmi, qui est elle-même un sous-groupe des langues soudaniques centrales. La figure 22 illustre la localisation des cinq premiers groupes sous leur ancien regroupement en tant que langues oubanguiennes et la figure 23 montre la répartition des langues sara-bongo-bagirmi dans leur entièreté.

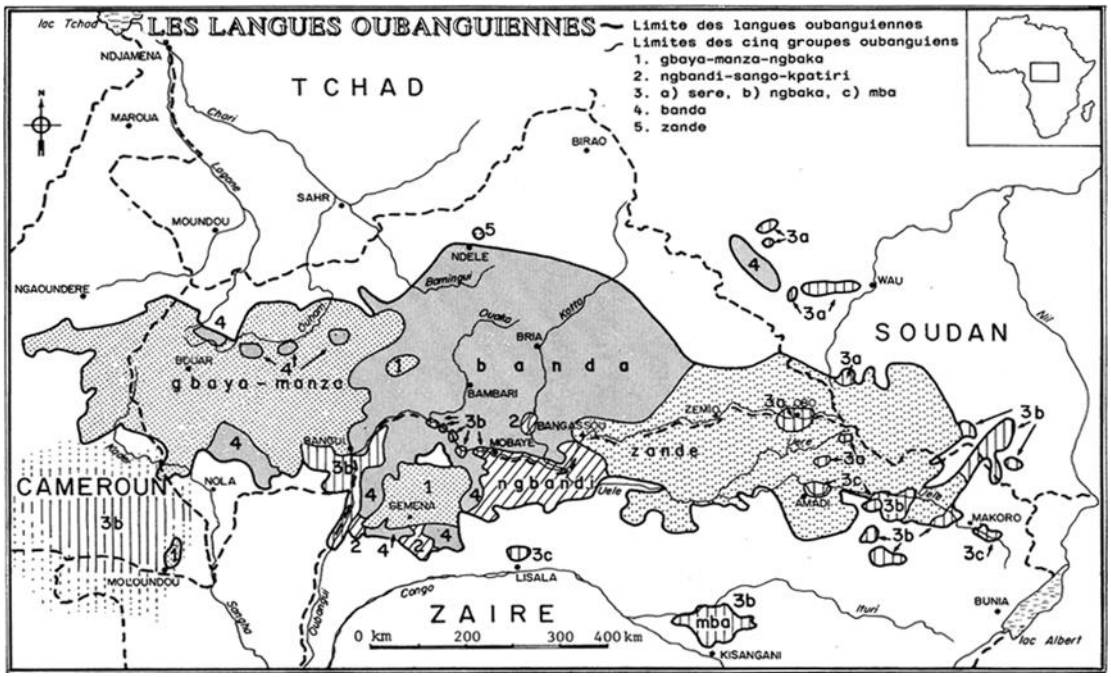


FIGURE 22. Les groupes oubanguiens (le groupe sere-ngbaka-mba, le groupe banda, le groupe ngbandi-mongoba-kazibati, le groupe zande) et le groupe gbaya-manza-ngbaka (anciennement aussi classés comme oubanguien) (Moñino 1988)

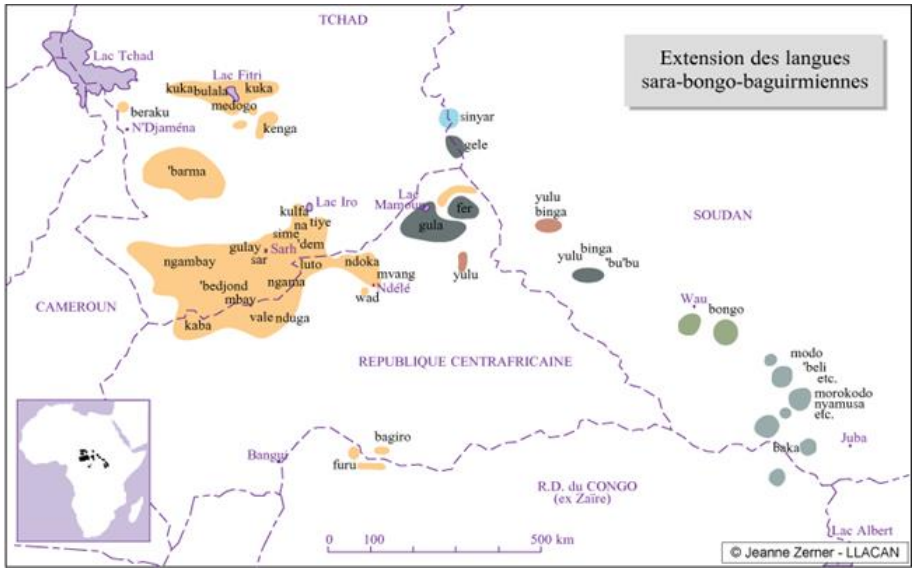


FIGURE 23. Les langues sara-bongo-bagirmi (Boyeldieu 2006a). Les langues sara-bongo-bagirmi occidentales sont en rose (par exemple, Kaba), gris foncé (par exemple, Gula ou Gele) et brun (par exemple, Yulu)

Le degré de diversité interne au sein des six groupes (les langues gbaya-manza-ngbaka, quatre groupes oubanguiens et les langues sara-bongo-bagirmi occidentales) est également assez faible. Certains de ces groupes auraient tout aussi bien pu être qualifiés de langues sans trop d'exagération. En outre, à l'intérieur de l'ensemble oubanguien, au moins sur la base de la similitude lexicale (cf. Boyeldieu & Cloarec-Heiss 1986; Moñino 1988:19), on peut dire que les langues sere-ngbaka-mba et les langues banda sont assez étroitement liées et forment probablement un groupe avec les langues ngbandi-mongoba-kazibati, le zande étant le seul à ne pas être apparenté de manière transparente aux trois autres groupes oubanguiens.

Un autre point important est qu'en plus d'être plutôt de faible profondeur et d'avoir un faible niveau de diversité interne, la plupart de ces groupes, si ce n'est tous, sont très probablement arrivés dans cette région d'Afrique centrale relativement récemment et ce n'est que lors de leur entrée dans cette région que la plupart des événements de spéciation au sein de ces groupes se sont produits. Par exemple, la figure 24 montre une reconstruction des routes migratoires des populations sara-bongo-bagirmi, indiquant le foyer d'origine du proto-sara-bongo-bagirmi dans ce qui est aujourd'hui le Soudan du Sud, juste en dehors de la région en question, le nœud de diversification du proto-sara-bongo-bagirmi occidental dans le nord-est de la République centrafricaine, à l'intérieur de la région en question, et un nœud important de diversification supplémentaire au sein des langues sara-bongo-bagirmi occidentales autour de la frontière entre le Tchad et la République centrafricaine, à savoir le nœud proto-sara. De même, comme on peut le voir sur la figure 22, le groupe sere des langues sere-ngbaka-mba et une partie des langues banda sont encore parlés au Soudan du Sud, approximativement dans la même zone que le foyer d'origine du proto-sara-bongo-bagirmi de la figure 24. En outre, les données disponibles suggèrent que la partie restante des langues banda et le groupe ngbaka-mba des langues sere-ngbaka-mba sont également venus dans la région en question à partir d'approximativement la même zone au Soudan du Sud (par exemple,

Rombi & Thomas 2006:22 pour les langues ngbaka-mba; Tisserant 1930:8–10 pour les langues banda).

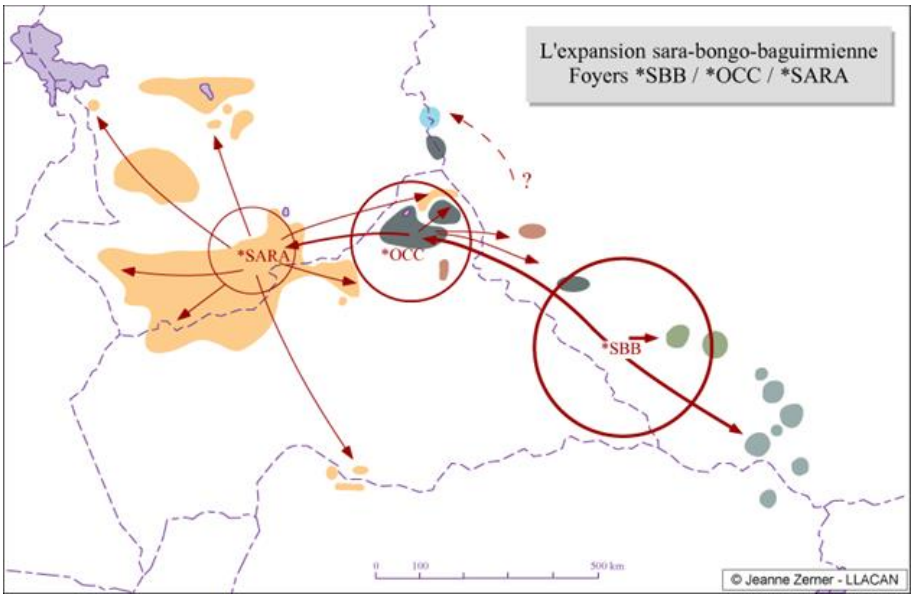


FIGURE 24. Une reconstruction des routes migratoires des populations sara-bongo-bagirmi (Boyeldieu 2006a) (*SBB est le foyer d'origine présumé du proto-sara-bongo-bagirmi ; *OCC est le nœud de diversification du proto-sara-bongo-bagirmi occidental ; *SARA est le nœud de diversification du proto-sara, un sous-groupe majeur au sein des langues sara-bongo-bagirmi occidentales).

En résumé, la région en question semble avoir été occupée relativement récemment par six groupes linguistiques de plutôt faible profondeur et plutôt homogènes et, parmi ces six groupes, au moins trois sont apparentés de manière transparente les uns aux autres et pourraient également être apparentés de manière plus éloignée à un autre groupe parmi ces six, ce qui réduit à trois le nombre total de groupes linguistiques différents impliqués. Ces groupes se sont installés dans la région en venant de l'extérieur. De plus, à l'exception d'un seul groupe, tous ont très probablement migré à partir de la même région générale dans ce qui est aujourd'hui le Soudan du Sud, ce qui rend probable qu'ils aient été en contact étroit les uns avec les autres avant même d'entrer dans la région en question. La seule exception probable est le groupe gbaya-manza-ngbaka, qui est plus probablement entré dans la région par le nord, quelque part dans le sud du Tchad et plus près de la majorité des langues restantes de l'Aire de Convergence Centrale. Ces groupes ont ensuite subi une diversification supplémentaire lors de leur entrée dans la région. Comme la proto-langue du groupe gbaya-manza-ngbaka possédait probablement déjà des MNFPs, nous nous retrouvons avec une très forte probabilité que les deux (ou au maximum trois) autres proto-langues issues de la même région aient simplement eu des MNFPs dès le départ ou les aient acquis en entrant dans la région. Dans une telle situation, il est facile d'imaginer comment la région concernée a pu facilement devenir aussi homogène qu'elle l'est en ce qui concerne la caractéristique en question en raison des effets fondateurs. En fait, les langues parlées

dans la région en question sont plutôt homogènes en ce qui concerne un certain nombre de traits qui sont par ailleurs très rares d'un point de vue typologique, comme la fréquence lexicale élevée des occlusives labiales-vélaires (cf. §5), la présence proéminente des battues labiales (Olson & Hajek 2003) et l'utilisation de constructions aux qualifiants exprimés comme des possédés (également connues sous le nom d'inversion de dépendance) (Van de Velde 2012; 2013:233–234).

Au vu des arguments présentés ci-dessus, il est extrêmement improbable que la présence importante des MNFPs canoniques dans cette région de l'Afrique centrale atteste de quelque manière que ce soit du rôle historique hypothétique de cette région en tant que foyer historique potentiel de l'Aire de Convergence Centrale. En même temps, comme nous le verrons dans §4.5.3, cette région d'Afrique centrale, avec sa présence importante des MNFPs canoniques, a dû servir de source pour la diffusion du trait MNFP parmi les langues bantu plus au sud, dans le corridor du fleuve Congo et au nord de la République démocratique du Congo. Compte tenu de l'orientation générale de l'Aire de Convergence Centrale et de sa dynamique des mouvements des populations, déterminée par l'écologie et la géographie de la région, il est très probable que le foyer historique primaire de l'Aire de Convergence Centrale se situe immédiatement au nord-ouest de la République centrafricaine, le long du corridor du fleuve Benue qui va du sud du Tchad au centre du Nigeria en passant par le nord du Cameroun. Il s'agit essentiellement de la région saillante sur le graphique d'intensité spatiale des langues avec des MNFPs en Afrique de la figure 15. La région située le long du corridor du fleuve Benue est densément peuplée, tant en termes de population (cf. la figure 21) que de langues (cf. la figure 13). De plus, le paysage linguistique de cette région est très fragmenté et caractérisé par une grande diversité linguistique, ce qui forme un contraste frappant avec la partie de l'Aire de Convergence Centrale en République centrafricaine plus au sud-est dont j'ai parlé en premier.

4.5.3 Des MNFPs optionnelles et/ou soumises aux restrictions sur l'utilisation : grammaticalement non canoniques et aréalement périphériques

Du point de vue de la dynamique du changement linguistique, l'optionalité des MNFPs et les restrictions sur leur utilisation en fonction de construction sont plus typiques des marques soit innovantes, soit en voie de disparition, selon la direction du changement. Des études comparatives détaillées seraient nécessaires pour déterminer cette direction avec certitude. Cependant, compte tenu de ce que je sais des langues en question, ma

forte impression est que ces MNFPs sont plus souvent des innovations que des rétentions de stades plus anciens en voie de disparition.

D'un point de vue de la typologie aréale, on s'attend à ce que les langues comportant des MNFPs facultatives et soumises aux restrictions sur l'utilisation en fonction de construction, c'est-à-dire les langues codées avec des valeurs pseudo-numériques 1, 2 et 3 dans le tableau 1 (que j'appellerai les types 1, 2, 3), soient situées principalement à la périphérie de l'aire des langues comportant des MNFPs. Cette prévision se confirme effectivement, comme on peut le constater en comparant la distribution spatiale de ces langues dans la figure 25 avec celle de toutes les langues ayant des MNFPs dans l'échantillon dans la figure 13 (cf. §4.3). Il est remarquable de constater que les langues de type 2 et 3 ont une répartition spatiale presque complémentaire, avec des chevauchements notables uniquement dans le corridor du fleuve Congo (cf. §4.3) et dans le sud-ouest du Cameroun, qui sont également les deux régions où l'on trouve les langues de type 1. De plus, les langues de type 2 sont presque toutes des langues bantoïdes et, au sein de ce groupe, elles sont presque toutes des langues bantu au sens strict (Narrow Bantu), alors que les langues de type 3 sont beaucoup plus diversifiées génétiquement, ce qui suggère que les langues bantoïdes et surtout les langues bantu étroites se distinguent par quelque chose dans leurs structures. D'un point de vue typologique, les langues bantu sont en effet connues pour être différentes à bien des égards des langues des régions plus septentrionales de l'Afrique subsaharienne, auxquelles elles sont toutefois apparentées (cf. Clements & Rialland 2008; Güldemann 2008).

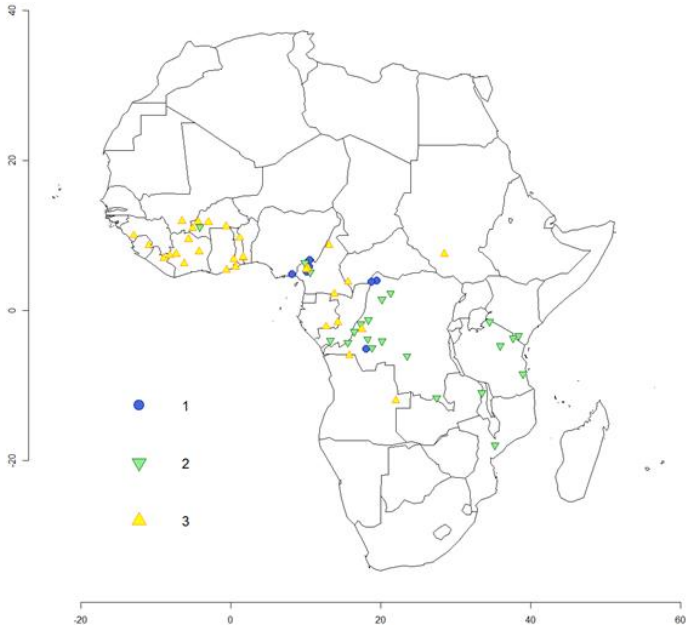


FIGURE 25. Langues avec des MNFPs optionnelles et/ou soumises aux restrictions sur l'utilisation en fonction de construction. Voir le tableau 1 (§4.4) pour la signification des valeurs pseudo-numériques : 1 = « optionnel » et « avec restrictions », 2 = « optionnel » et « sans restrictions », 3 = « obligatoire » et « avec restrictions ».

Une explication possible de la distribution observée du type 2 est que les MNFPs dans les langues bantoïdes et surtout dans les langues bantu ont tendance à se développer par des voies un peu différentes qu'ailleurs, les voies qui sont moins susceptibles de conduire à des restrictions sur l'utilisation en fonction de construction mais qui sont plus susceptibles d'aboutir à des MNFPs optionnelles, soit dans le sens qu'elles sont en fin de phrase mais optionnelles dans cette position, soit que leur emplacement en fin de phrase est optionnel. Comme le décrivent Devos & van der Auwera (2013), « recurrent sources for post-verbal negative markers [y compris les MNFPs] in Bantu languages are locative pronouns, possessive pronouns and negative (answer) particles », qui semblent en effet être rarement attestées comme sources de MNFPs dans les régions plus septentrionales de l'Afrique subsaharienne. Comme mentionné à propos de la marque de négation *bɔ* dans la langue bantu nzadi dans §4.2.5, qui est optionnellement en fin de phrase, les pronoms possessifs comme source de marques de négation post-verbales sont, par exemple, peu susceptibles d'avoir pour position d'origine la position en fin de phrase. Dans une certaine mesure, il en va de même pour les pronoms locatifs dans les langues bantu. Les particules de réponse négative en tant que source de marques de négation post-verbales, bien que susceptibles d'avoir pour position d'origine la position en fin de phrase, sont peu susceptibles d'être soumises à des restrictions sur l'utilisation en fonction de construction.

L’optionalité fréquente des MNFPs en bantu doit avoir beaucoup à voir avec leur âge relativement jeune. Le caractère innovant relativement récent des MNFPs en bantu est confirmé par leur distribution restreinte au sein du bantu et par la variation importante de leurs formes à travers le bantu, ce qui contraste fortement avec l’uniformité relative et la présence obligatoire presque universelle des marques de négation préverbales plus anciennes (cf. Kamba Muzenga 1981; Güldemann 1999; Devos & van der Auwera 2013). De plus, alors que les formes des anciennes marques de négation pré-verbales peuvent être reconstruites jusqu’au proto-bantu, on ne peut leur fournir une étymologie autre qu’une marque de négation (Kamba Muzenga 1981). En même temps, les MNFPs ne peuvent pas être reconstruits jusqu’au proto-bantu et, lorsque leur étymologie peut être établie, elles proviennent souvent d’éléments qui ne sont pas des marques de négation. La seule exception notable est celle des MNFPs qui proviennent de particules de réponse négatives.

Un autre facteur important contribuant à l’optionalité fréquente des MNFPs en bantu est que, typiquement, les marques de fin de phrase en tant que telles ne sont pas un trait morphosyntaxique proéminent des langues bantu et, à cet égard, elles diffèrent clairement des langues du nord de l’Afrique subsaharienne (cf. Idiatov 2012a). Alors que la présence proéminente des marques de fin de phrase dans la morphosyntaxe des langues de l’Afrique subsaharienne septentrionale serait propice à la mise à niveau des MNFPs innovantes d’un statut optionnel à un statut obligatoire, un tel facteur d’attraction fait généralement défaut dans les langues bantu.

Au sein du bantu, le corridor du fleuve Congo est clairement la zone de concentration de l’innovation des MNFPs. C’est ce que suggère déjà l’observation de la figure 25, qui montre que les langues bantu de type 3 sont essentiellement confinées à cette région et que la concentration des langues bantu de type 2 est également la plus élevée dans cette même région. L’importance du corridor du fleuve Congo devient encore plus évidente lorsque nous incluons les langues bantu de type 4, c’est-à-dire les langues bantu avec des MNFPs obligatoires et sans restriction sur l’utilisation en fonction de construction. Ainsi, comme on peut le voir sur la figure 26, les langues bantu ayant un plus grand nombre de MNFPs canoniques sont également concentrées dans le corridor du fleuve Congo.

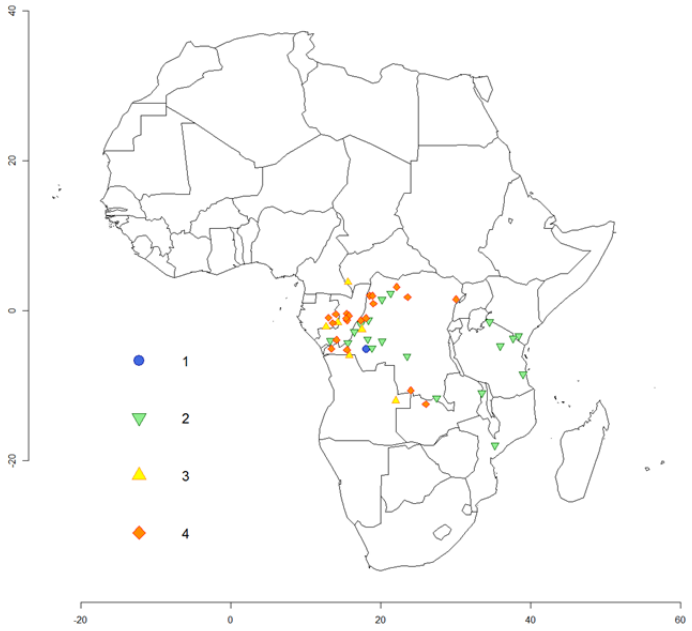


FIGURE 26. Langues bantu avec MNFPs. Voir le tableau 1 (§4.4) pour la signification des valeurs pseudo-numériques : 1 = « optionnel » et « avec restrictions », 2 = « optionnel » et « sans restrictions », 3 = « obligatoire » et « avec restrictions », 4 = « obligatoire » et « sans restrictions ».

En outre, nous pouvons observer dans la figure 26 deux faibles bandes de langues bantu avec des MNFPs qui semblent être liées aux extrémités nord et sud du corridor du fleuve Congo et qui vont toutes deux vers le sud-est, l'une au nord et l'autre au sud du bassin du fleuve Congo. De ces deux zones de proéminence secondaire, celle du sud est clairement une ramification historique du couloir du fleuve Congo, tandis que celle du nord doit partager son origine avec le couloir du fleuve Congo dans l'Aire de Convergence Centrale, située plus au nord en République centrafricaine (cf. la figure 19 dans §4.5.1.2), comme l'illustre schématiquement la figure 27. Certes, nous ne pouvons pas totalement exclure la possibilité que les MNFPs optionnelles des langues bantu d'Afrique de l'Est aient évolué de manière indépendante.

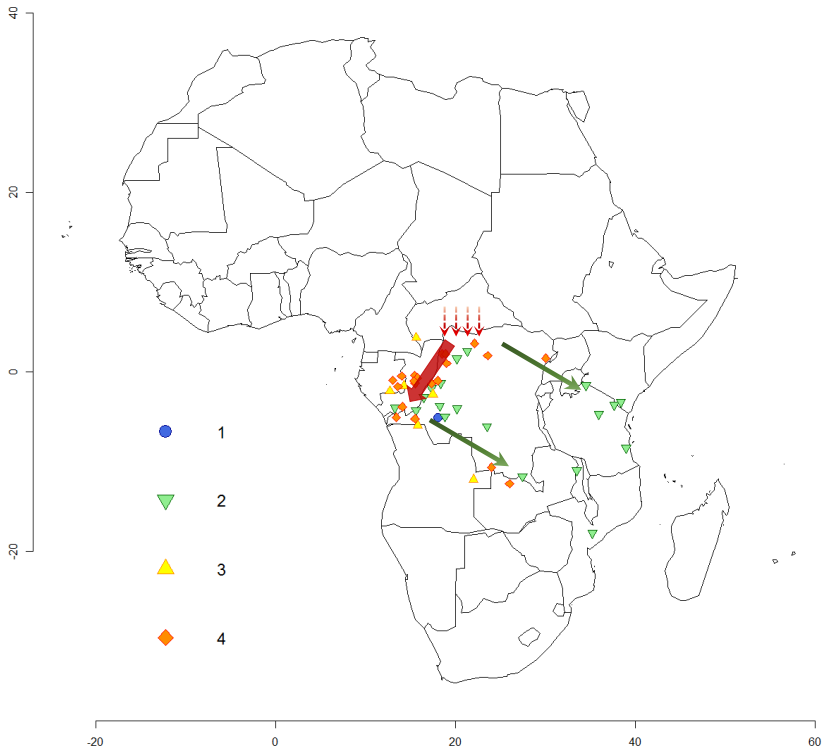


FIGURE 27. La direction suggérée de la propagation de l'utilisation des MNFPs au sein des langues bantu à partir de l'Aire de Convergence Centrale dans le nord de l'Afrique centrale vers le couloir du fleuve Congo et les deux zones de proéminence secondaires.

L'émergence du corridor du fleuve Congo et des deux zones de proéminence secondaire doit résulter de mouvements de population et/ou de langues relativement récents au départ de l'Afrique centrale. Il est clair qu'ils ont eu lieu bien plus tard que l'expansion bantu initiale à l'intérieur et autour du bassin du fleuve Congo. À cet égard, il faut comparer le profil aréal de la distribution des MNFPs en bantu avec la route d'expansion des langues bantu reconstruite par Grollemund et al. (2015) et reproduite dans la figure 28. La comparaison montre clairement que l'expansion vers le sud-ouest de l'utilisation des MNFPs dans le corridor du fleuve Congo s'est faite dans la direction opposée à la route originale de l'expansion bantu dans la moitié nord du bassin du fleuve Congo. La zone de proéminence secondaire nord ne correspond à aucune route originelle de l'expansion bantu dans cette région à partir du nord de la République démocratique du Congo. La zone de proéminence secondaire sud correspond partiellement à une route originelle d'expansion bantu dans cette région, mais elle n'a pas pu se former avant l'émergence de la zone du corridor du fleuve Congo et n'a donc pas pu coïncider dans le temps avec cette partie de l'itinéraire d'expansion bantu.

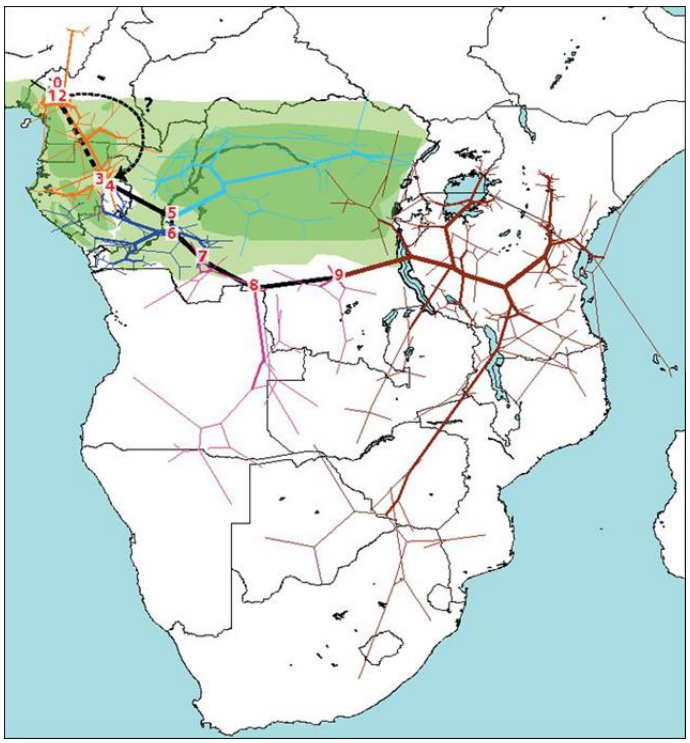


FIGURE 28. Route migratoire bantu reconstruite par Grollemund et al. (2015) sur l'arbre de consensus en utilisant les emplacements géographiques des langues contemporaines et en reliant les emplacements ancestraux par des lignes droites (la véritable route sera différente) (les positions numérotées correspondent aux principaux nœuds de diversification sur l'arbre de consensus ; la ligne pointillée courbe indique la route migratoire suggérée à travers les corridors de savane ; l'ombrage plus clair (vert) correspond à la délimitation de la forêt tropicale à 5 000 ans avant le présent ; l'ombrage plus foncé (vert) correspond à la délimitation de la forêt tropicale à 2 500 ans avant le présent).

4.6 Observations finales

Les distributions synchroniques que nous découvrons sont nécessairement le produit de l'évolution de la langue dans le temps et l'espace. J'ai proposé ici une analyse de la dynamique spatio-temporelle des langues en Afrique subsaharienne en ce qui concerne le trait MNFP. Une question que l'on est souvent tenté de poser lorsqu'on fait de la typologie aréale est de savoir quel groupe linguistique, parmi ceux présents dans la région où le trait est proéminent, aurait pu être le vecteur primaire de ce trait. Or, cette question ne peut avoir une réponse pleinement significative que si l'on sait qu'aucun groupe linguistique n'a disparu de la scène sans laisser de traces depuis l'émergence du trait dans la région. Malheureusement, dans une région comme l'Afrique subsaharienne, nous ne pouvons pas en être sûrs. En fait, nous pouvons être tout à fait sûrs du

contraire.[21] De plus, le fait que tous les membres d'un certain groupe linguistique portent le trait et sont parlés à l'intérieur de la région où le trait est bien présent ne peut pas prouver que ce groupe linguistique est le vecteur principal du trait.

Par exemple, Dryer (2009:346) envisage deux scénarios possibles concernant sa région principale des langues VO&VNeg en Afrique centrale, qui, sans surprise, coïncide largement avec notre Aire de Convergence Centrale du trait MNFP.[22] Le premier scénario est que le trait VO&VNeg provient des langues tchadiques, parce que ce trait est omniprésent dans les langues tchadiques et que toutes les langues tchadiques sont parlées dans la région concernée. Cependant, les langues tchadiques sont aussi typologiquement très différentes de leurs parents afro-asiatiques éloignés, dont la plupart ne portent pas non plus le trait en question. En d'autres termes, nous ne pouvons pas savoir si le trait peut être reconstruit au proto-tchadique indépendamment de la participation actuelle des langues tchadiques dans l'aire en question. Le second scénario est que la présence de la caractéristique en tchadique est due à une influence de substrat provenant des langues dites nilo-sahariennes et que cette influence de substrat pourrait avoir un effet si profond sur les langues tchadiques en raison de la taille relativement petite de la région peuplée par les locuteurs des langues tchadiques. Le groupe nilo-saharien qui est à la fois parlé dans la même région que les langues tchadiques et classé positif pour le trait pertinent par Dryer (2009:311) est le sara-bongo-bagirmi occidental. Cependant, comme indiqué dans §4.5.2, les langues sara-bongo-bagirmi occidentales font partie des nouveaux arrivants dans la région. En d'autres termes, même si le tchadique a été influencé par un substrat, ce qui est en fait tout à fait plausible étant donné les différences typologiques entre le tchadique et ses parents afro-asiatiques, nous ne pouvons pas savoir quel était ce substrat, ni si l'omniprésence du trait en tchadique peut être attribuée à ce substrat.

Les considérations ci-dessus s'appliquent également au rôle du tchadique et d'autres groupes homogènes similaires (comme le gbaya-manza-ngbaka et le zande) dans notre Aire de Convergence Centrale du trait MNFP. Alors qu'une distribution homogène du trait au sein d'un groupe linguistique donné (tous les membres du groupe portent le trait et sont parlés à l'intérieur de la région) ne nous renseigne pas beaucoup

[21] Par exemple, voir Kleinewillinghöfer (2001) sur le jalaa, un isolat linguistique apparent dans le nord-est du Nigeria, la région qui est particulièrement pertinente pour le trait MNFP, qui a disparu assez récemment et pour laquelle seules quelques données lexicales ont pu être collectées auprès de locuteurs qui avait juste des souvenirs de la langue.

[22] Ce n'est pas très surprenant car Dryer (2009) limite son étude aux marques de négation qui sont des mots. Bien que tous les marques de négation post-verbales qui sont des mots ne soient pas également en fin de phrase, les MNFPs sont presque toujours analysées comme des mots précisément en raison de leur orientation phrastique et, par conséquent, au moins les MNFPs canoniques seraient toujours classées comme des marques de négation post-verbales dans la typologie de Dryer (2009).

sur la dynamique spatio-temporelle du trait, un signal beaucoup plus informatif est généralement fourni par des groupes qui sont diversifiés en ce qui concerne le trait, avec des membres à l'intérieur et à l'extérieur de la région, surtout lorsqu'il peut être complété par des informations indépendantes sur les mouvements des populations et des langues. Ainsi, dans le cas du trait MNFP, divers groupes niger-congo sont parlés autour du corridor du fleuve Benue dans l'Aire de Convergence Centrale, ainsi que plus à l'ouest dans l'Aire de Convergence Occidentale. En même temps, de nombreux groupes niger-congo sont également parlés en dehors de l'Aire de Convergence Centrale et de l'Aire de Convergence Occidentale. Nous pouvons raisonnablement supposer que ces derniers groupes niger-congo ont perdu le trait MNFP lorsqu'ils se sont déplacés en dehors de l'aire, comme lorsqu'ils ont pénétré dans la zone forestière le long de la côte du golfe de Guinée, par exemple, en raison de l'influence d'un substrat (tout comme, par exemple, ils ont développé une fréquence lexicale élevée de labiales-vélaires dans les mêmes régions côtières comme je démontre dans §5), ou lorsque des groupes dépourvus de ce trait ont pénétré dans la région depuis l'extérieur. Par exemple, comme nous l'avons vu dans §4.5.1.1, le premier type de perte est susceptible d'être le cas pour les langues yoruboïdes dans la brèche côtière du sud-ouest du Nigeria entre deux extensions côtières de l'Aire de Convergence Centrale, tandis que le second type de perte est susceptible d'être la raison de l'émergence de la discontinuité majeure séparant l'Aire de Convergence Centrale de l'Aire de Convergence Occidentale. En même temps, il est plutôt improbable que le trait MNFP soit reconstruit à des nœuds plus élevés dans l'arbre niger-congo, pas même au nœud proto-benue-congo. Un contre-argument majeur à une reconstruction aussi profonde est présenté par l'absence générale des MNFPs dans les langues bantoïdes du sud (avec l'exception la plus notable des langues bantu du couloir du fleuve Congo mais, comme discuté dans §4.5.3, il s'agit clairement d'un développement récent). Au sein des langues benue-congo, leur groupe superordonné, et au sein des langues niger-congo en général, les langues bantoïdes du sud et surtout les langues bantu sont généralement considérées comme archaïques dans leur profil typologique (par exemple, Hyman 2011).

5 La fréquence lexicale des occlusives labiales-vélaires dans le nord de l'Afrique subsaharienne

5.1 Introduction

Les occlusives labiales-vélaires (LV), telles que *k͡p*, *g͡b* et *ŋ͡m*, sont des sons qui sont produits avec des gestes presque simultanés de fermeture vélaire et labiale (Ladefoged & Maddieson 1996:332–343). Les occlusives LV sont présentes dans de nombreuses langues de l'ouest et du centre de l'Afrique subsaharienne septentrionale, alors qu'elles sont rares ailleurs (Cahill 2008; 2018; Maddieson 2011; 2018). En mettant en évidence que l'ensemble des langues à occlusives LV est géographiquement cohérent mais généalogiquement divers, Güldemann (2008:156–158) et Clements & Rialland (2008) utilisent leur présence comme l'une des traits définissant une aire linguistique, qu'ils appellent respectivement la « Macro-Sudan Belt » et la « Sudanic zone ». La prépondérance d'un tel trait typologiquement inhabituel dans une région étendue et généalogiquement diverse soulève la question de savoir où et comment il est apparu et par quel mécanisme il s'est répandu. Les hypothèses proposées dans la littérature s'appuient sur les outils explicatifs habituels de la linguistique aréale, tels que l'héritage, l'innovation par changement phonologique, l'emprunt de phonèmes par le biais des emprunts, l'interférence du substrat, et un concept plus abstrait et moins bien défini de diffusion. Ainsi, Westermann (1911; 1927) a suggéré que le trait est une innovation par le changement phonologique des occlusives vélaires labialisées aux occlusives LV. En appliquant la règle de la majorité, Greenberg (Greenberg 1983:8–9) a suggéré que le trait devrait être reconstruit au proto-niger-congo et qu'il a été conservé dans certaines des langues filles et diffusé par le biais des emprunts ou des changements phonologiques convergents (« convergent sound change ») dont le vecteur était vraisemblablement des locuteurs bilingues ramenant des occlusives LV dans leurs communautés linguistiques primaires. La même logique de la règle de la majorité est appliquée par Cahill (2017; 2018), qui plaide généralement en faveur de l'héritage à partir de la proto-langue non seulement pour le niger-congo et ses branches principales, mais aussi dans d'autres groupes, tels que le soudanique central. Un article non publié de Vogler (2014) s'oppose à la plausibilité de l'argument majoritaire en faveur de l'hypothèse de l'héritage et suggère une origine du trait à partir d'un substrat inconnu dans la partie occidentale de

la forêt tropicale ouest-africaine (la zone centrée sur le domaine moderne des langues kru) d'où il s'est diffusée ailleurs.

Afin d'essayer de répondre aux questions du où et du comment de l'origine et de la propagation des occlusives LV, je propose d'aller au-delà de la pratique habituelle de la linguistique aréale qui consiste à examiner la distribution géographique d'un trait en termes de présence ou d'absence dans des langues données. Suivant cette nouvelle méthodologie, j'étudie la profondeur de l'ancrage des occlusives LV dans les langues où elles sont attestées en estimant leurs fréquences lexicales dans une grande base de données lexicales et j'analyse la distribution spatiale de ces fréquences. Je compare également ces résultats à des données provenant d'un certain nombre d'autres domaines pertinents, tels que la géographie, la paléoanthropologie, l'anthropologie moléculaire et l'anthropologie culturelle, l'archéologie, la paléoclimatologie et la paléobotanique. En exploitant la variation des données au lieu d'essayer de les réduire, je parviens à un scénario historique beaucoup plus riche et détaillé que ne l'aurait permis l'approche réductionniste traditionnelle.

Je pars de l'observation que les langues individuelles qui ont des occlusives LV dans leur inventaire de phonèmes diffèrent grandement en termes de proéminence de ces consonnes. Les occlusives LV sont des phonèmes plutôt marginaux dans la plupart des langues de l'Afrique subsaharienne septentrionale que je connais, dans le sens où elles sont moins fréquentes que les autres occlusives et/ou limitées à des positions spécifiques dans le mot et/ou à des parties du vocabulaire. En revanche, on connaît aussi des langues comme le yoruba [yoru1245], où les occlusives LV sont des phonèmes consonantiques normaux. Principalement grâce à l'existence de la base de données RefLex (Segerer & Flavier 2011–2022), j'ai pu estimer la fréquence lexicale des occlusives LV dans 315 langues africaines qui les ont dans leur inventaire phonologique et comparer l'estimation pour chaque langue à une estimation de la situation canonique dans laquelle chaque phonème consonantique a une fréquence égale dans le lexique. J'ai ensuite étudié la distribution spatiale de l'intervalle entre les deux estimations. Les profils géographiques qui en ressortent se prêtent à une interprétation historique claire et intéressante. Je montre qu'il existe trois foyers de haute fréquence lexicale des occlusives LV, qui présentent toutes les caractéristiques des zones de refuge. Ils sont séparés les uns des autres par des zones à faible fréquence lexicale d'occlusives LV qui correspondent à des zones dont le climat et la végétation sont mieux adaptés aux migrants provenant d'un habitat de savane. Je conclue donc que la distribution aréale actuelle des occlusives LV est due à la rétention aréale de ces consonnes elles-mêmes et/ou d'autres caractéristiques phonétiques qui facilitent leur émergence et leur préservation une fois émergées.

Le reste du chapitre est organisé comme suit. §5.2 traite de la quantité, de la qualité et de l'origine des données (§5.2.1), explique comment j'ai estimé la fréquence lexicale des occlusives LV dans les langues africaines et présente les résultats de ces estimations (§5.2.2). Je teste également l'hypothèse selon laquelle les occlusives LV sont plus fréquentes dans les mots expressifs que dans le vocabulaire général, une hypothèse que j'ai déduite de l'examen des langues individuelles et qui est pertinente pour mon explication de l'origine et de la propagation des occlusives LV (§5.2.3). §5.3 présente la distribution spatiale des fréquences lexicales des occlusives LV en utilisant deux méthodes de visualisation différentes, l'interpolation spatiale (§5.3.1) et la modélisation additive généralisée (§5.3.2). Les deux méthodes convergent vers la même structure spatiale composée de trois foyers, dont deux ne sont séparés que par une étroite discontinuité. J'ai également procédé à une validation de ces résultats par recoupement en cartographiant les toponymes africains orthographiés avec des occlusives LV, ce qui produit trois groupes qui correspondent étroitement aux trois foyers (§5.3.3). Mon explication de l'origine et de la propagation des occlusives LV implique de manière cruciale le phénomène de la prosodie d'emphase-C qui facilite à la fois l'émergence des occlusives LV et leur transfert par contact linguistique, et explique la fréquence plus élevée des occlusives LV dans les parties les plus expressives du lexique. Je discute brièvement du phénomène de la prosodie d'emphase-C dans §5.4. Dans §5.5, je discute des implications historiques des résultats, à savoir que les occlusives LV, ou du moins les traits qui contribuent à leur émergence, étaient présentes dans et autour des foyers actuels avant l'arrivée des familles linguistiques qui y sont actuellement parlées (§5.5.1), mais qu'elles constituent une innovation dans toutes ou la plupart de ces familles linguistiques (§5.5.3). Cette préservation du trait dans certaine régions géographiques est principalement due à une conversion linguistique dans les foyers de haute fréquence lexicale et à des emprunts dans les autres régions. En comparant mes résultats avec des données provenant d'un certain nombre d'autres domaines pertinents, je propose un scénario détaillé pour l'émergence initiale et la propagation des occlusives LV et/ou des traits phonétiques qui facilitent leur émergence au niveau macro de l'Afrique subsaharienne septentrionale (§5.5.1). Mes résultats me permettent également d'ajuster et d'affiner les scénarios proposés dans la littérature pour l'expansion bantu, l'un des plus grands événements d'expansion linguistique de l'histoire humaine récente (§5.5.2). Enfin, la distribution géographique des fréquences lexicales élevées des occlusives LV suggère que les occlusives LV ne devraient pas être reconstruites dans les proto-langues des principales sous-branches du niger-congo et du soudanique central, à moins qu'elles n'aient été parlées dans l'un des foyers, ni dans le proto-niger-congo et le proto-soudanique central (§5.5.3).

5.2 Estimation de la fréquence des occlusives labiales-vélaires dans le lexique

5.2.1 La base des données

La principale source de données que j'ai utilisée est RefLex (Segerer & Flavier 2011–2022), une base de données en ligne de plus d'un millier de lexiques de langues africaines, qui est accompagnée d'un certain nombre d'outils utiles pour la reconstruction et l'analyse statistique. L'exactitude de son contenu est facilement vérifiable grâce aux liens intégrés vers ses sources publiées. J'ai laissé de côté les sources publiées avant 1900, car la notation des occlusives LV n'est pas fiable dans beaucoup d'entre elles. De plus, je n'ai pas tenu compte des sources comportant moins de cent entrées. Ce seuil, en partie arbitraire, a été choisi pour inclure des sources suffisamment importantes pour être un minimum représentatives d'une langue, sans pour autant exclure les listes de mots de vocabulaire de base qui sont les seules sources d'information existantes sur de nombreuses langues africaines.[23] Lorsqu'une même langue est représentée par plus d'une source dans RefLex, j'ai choisi la meilleure source en termes de taille et de fiabilité. De plus, j'ai ajouté certaines sources lexicales des langues bantu et mande qui ne sont pas (encore) intégrées dans RefLex et j'ai utilisé les informations sur la présence ou l'absence des occlusives LV dans les inventaires phonologiques des langues africaines qui sont disponibles dans Phoible (Moran, McCloy & Wright 2014; Moran & McCloy 2019).

Mon échantillon contient 1110 langues, dont 545 ont des occlusives LV dans leur inventaire et 565 n'en ont pas. Je dispose de données sur la fréquence lexicale des occlusives LV pour 315 des 545 langues qui en possèdent. La figure 29 représente les langues de l'échantillon et met en évidence les zones de leur concentration dans l'espace, c'est-à-dire elle montre leur intensité spatiale. Il existe deux zones à forte concentration

[23] Sur la base d'une étude expérimentale utilisant des données provenant de langues australiennes, Dockum & Bowern (2019:50) soutiennent qu'on risque très probablement de représenter de manière incorrecte des faits de base concernant la phonologie de la langue quand on utilise des listes de mots inférieures au seuil de 400 mots. Dans cette perspective, même notre seuil de 100 mots est vraiment une solution du pauvre. Néanmoins, pour les besoins de ma recherche, il s'est avéré que l'inclusion de sources dont la taille est inférieure au seuil de 400 mots n'a pas affecté les résultats de manière significative, et ce malgré le fait que ces sources constituent environ 40% de mon échantillon de langues avec des données sur la fréquence lexicale des occlusives LV (cf. §5.3.2.2). Ainsi, en utilisant uniquement les sources de 400 mots et plus, j'ai obtenu des résultats très similaires à ceux de l'échantillon dans son ensemble.

de langues, toutes deux situées dans la ceinture dite de fragmentation linguistique. La plus importante est située à la frontière entre le Cameroun et le Nigeria. La seconde couvre une grande partie de l'Afrique de l'Est et se concentre autour du lac Victoria.

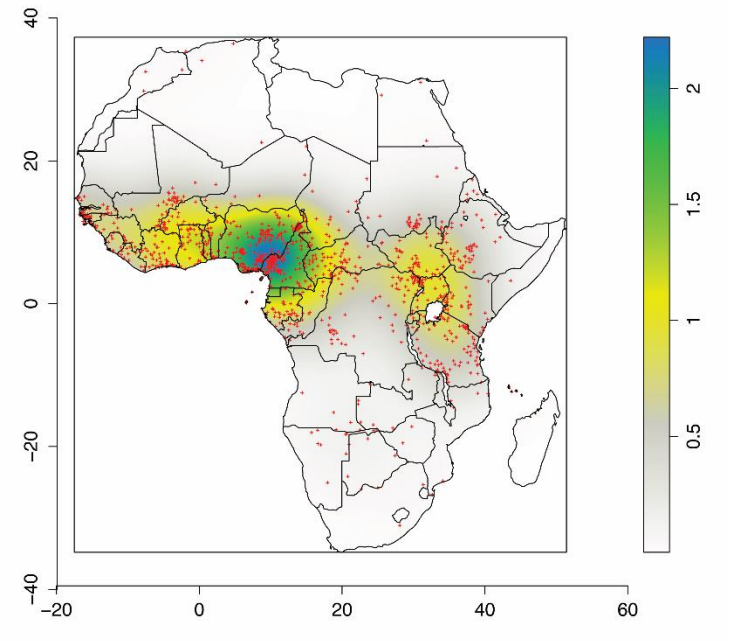


FIGURE 29. La distribution géographique des 1110 langues de l'échantillon et leur intensité spatiale (estimée par un lissage à noyau gaussien dont la largeur de bande a été optimisée en utilisant l'erreur quadratique moyenne).

Les deux figures suivantes montrent la distribution géographique des langues avec (figure 30) et sans (figure 31) occlusives LV ainsi que leur intensité spatiale. Comme prévu, la distribution des langues avec occlusives LV présentée dans la figure 30 correspond à l'aire Macro-Sudan Belt de Güldemann (2008) et à la Sudanic zone de Clements & Rialland (2008). De plus, la figure 30 montre que les langues à occlusives LV sont particulièrement concentrées dans une zone orientée est-ouest au sud du Nigeria. La figure 31 montre des zones de concentration de langues sans occlusives LV à l'extrême ouest et à l'est de l'Afrique subsaharienne septentrionale avec une bande d'intensité spatiale de ces langues un peu plus faible s'étendant d'ouest en est à travers la partie occidentale de l'Afrique subsaharienne septentrionale jusqu'au nord de la zone avec de nombreuses langues à occlusives LV plus proches de la côte. On observe également une forte concentration de langues sans occlusives LV dans une zone s'étendant du nord-est du Nigeria à l'ouest du Cameroun, qui recouvre partiellement la zone de plus forte concentration de langues à occlusives LV. Ce chevauchement partiel est dû à la forte fragmentation linguistique de cette zone. Notons toutefois que l'orientation spatiale des deux zones de chevauchement est différente. La zone à occlusives LV est située plus à l'ouest et a une orientation horizontale, alors que la zone sans occlusives LV a une orientation nord-sud. Ces configurations réapparaîtront dans

§5.3, où j'examine la distribution spatiale de la fréquence lexicale des occlusives LV dans les langues qui ont ces consonnes dans leur inventaire de phonèmes.

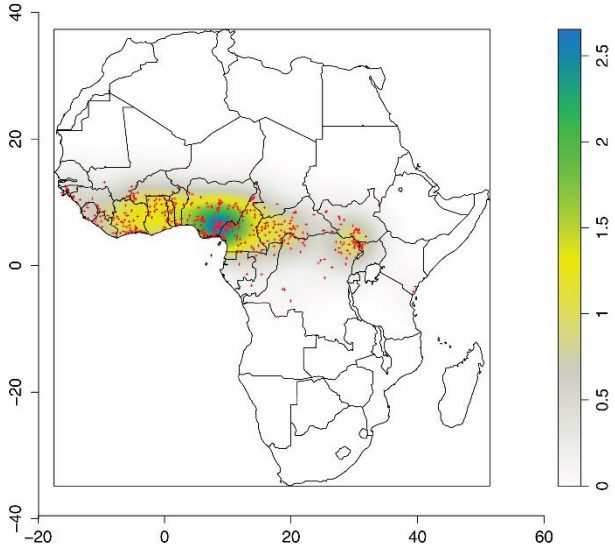


FIGURE 30. La distribution géographique des 545 langues avec occlusives LV et leur intensité spatiale (estimée par un lissage à noyau gaussien dont la largeur de bande a été optimisée en utilisant l'erreur quadratique moyenne).

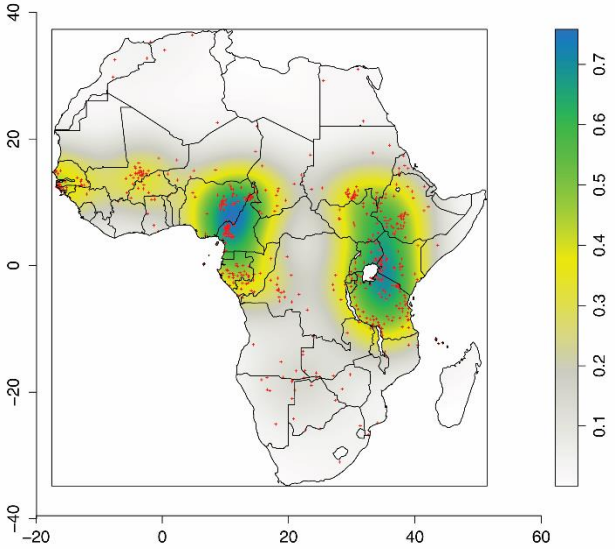


FIGURE 31. La distribution géographique des 565 langues sans occlusives LV et leur intensité spatiale (estimée par un lissage à noyau gaussien dont la largeur de bande a été optimisée en utilisant l'erreur quadratique moyenne).

5.2.2 Estimation de la fréquence des occlusives labiales-vélaires dans le lexique

Nous nous tournons maintenant vers les 315 langues de ma base de données qui possèdent des occlusives LV et pour lesquelles je dispose d'informations lexicales. Afin d'estimer la fréquence des occlusives LV dans leur lexique, j'ai commencé par nettoyer les données récoltées dans RefLex afin de les normaliser et de supprimer les erreurs. J'ai supprimé certaines langues dans lesquelles les combinaisons de symboles *kp* et/ou *gb* sont utilisées pour les successions d'une occlusive vélaire et d'une occlusive bilabiale. Ensuite, j'ai recodé les digraphes non reconnus comme tels par RefLex. J'ai également corrigé certaines conventions orthographiques liées à la représentation des occlusives LV qui diffèrent de l'IPA, comme le yoruba [yoru1245] *p* correspondant à /k͡p/. Enfin, j'ai séparé les clusters que RefLex traite automatiquement comme des phonèmes complexes, tels que les occlusives dites prénasalisées (successions d'une occlusive nasale et orale homorganique), par exemple *nd*, *mb*, les consonnes suivies de marques de labialisation (par exemple *bw*) ou de palatalisation (par exemple *by*) et les successions d'une occlusive et d'une fricative labiodentale (par exemple *bv*).

La formule que j'ai utilisée pour estimer la fréquence lexicale des occlusives LV dans une langue est donnée dans (24). Elle exprime la fréquence des occlusives LV en pourcentage, de sorte que 0% correspond à l'absence d'occlusives LV et 100% correspond au nombre d'occlusives LV qui serait attendu dans la situation canonique où tous les phonèmes consonantiques de la langue ont exactement la même probabilité d'occurrence dans le lexique. J'appellerai cette fréquence (i.e. F_{LV} = 100%) la *fréquence de référence*.[24]

(24) $$F_{LV} = \frac{LV_O}{LV_E} * 100\% = \frac{\sum T_{LV}}{\frac{\sum T_C}{\sum P_C} * \sum P_{LV}} * 100\%$$

Comme le montre (24), la fréquence estimée des occlusives LV dans une langue (F_{LV}) est le quotient du nombre observé d'occlusives LV dans ma source lexicale pour cette langue (LV_O) et du nombre attendu d'occlusives LV dans la situation canonique où chaque phonème consonantique est de fréquence égale dans le lexique (LV_E). Il va sans dire que le LV_E est un point de calibrage purement théorique et qu'il n'est en rien attendu qu'il existe. Il est calculé en divisant le nombre total de consonnes observées dans la source de données ($\sum T_C$) par le nombre total de phonèmes consonantiques de la langue ($\sum P_C$), puis en multipliant ce quotient par le nombre de consonnes occlusives LV dans l'inventaire des phonèmes de la langue ($\sum P_{LV}$).

Les résultats de l'estimation de la F_{LV} (en pourcentage) dans les 315 langues de l'échantillon qui ont des occlusives LV et pour lesquelles j'ai des données lexicales sont résumés dans la figure 32 au moyen d'un diagramme de densité de probabilité tronqué à zéro. Le diagramme de densité de probabilité est superposé aux mêmes données présentées au moyen d'un histogramme. La figure 32 montre également la médiane de la distribution de F_{LV} et la fréquence de référence de 100%. La médiane est très

[24] Afin de faciliter la visualisation et l'analyse statistique, je considère la fréquence lexicale relative de l'ensemble des occlusives LV dans une source donnée plutôt que les fréquences lexicales de chaque occlusive LV séparément. En théorie, ce choix pourrait conduire à des résultats trompeurs s'il existait des zones où de nombreuses langues présentent un ensemble de deux ou plusieurs occlusives LV, dont l'une a une fréquence lexicale élevée, alors que les autres sont marginales. Cependant, une telle zone n'existe pas et il n'y a qu'une poignée de langues dans mon échantillon où une LV a une fréquence lexicale de plus de 66% de la fréquence de référence, alors que la fréquence des autres est inférieure à 33%.

inférieure à la fréquence de référence, ce qui montre que les occlusives LV sont des phonèmes relativement rares dans la grande majorité des langues qui en possèdent.

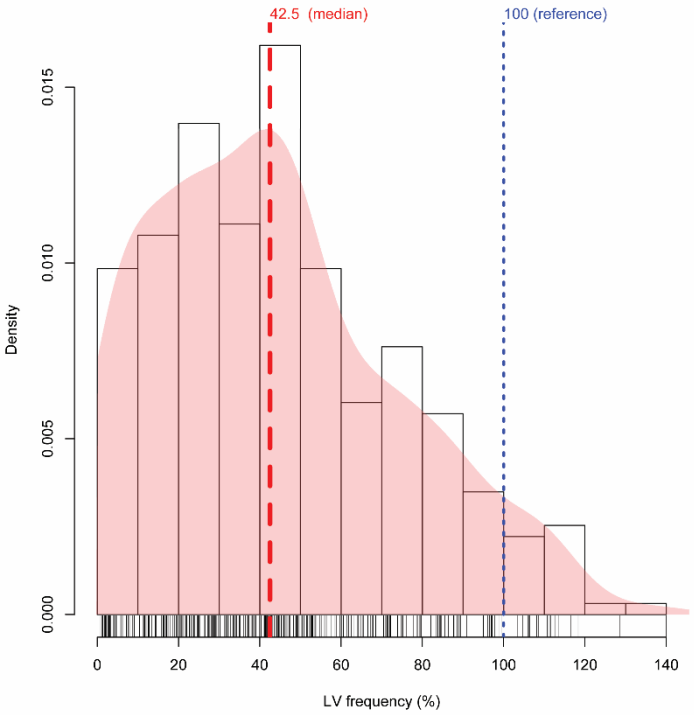


FIGURE 32. La densité de probabilité de toutes les fréquences F_{LV} dans mon échantillon en pourcentage (estimée par un lissage à noyau gaussien dont la largeur de bande a été optimisée à l'aide d'un sélecteur de largeur de bande à deux étapes ; les données les plus proches de zéro ont été pondérées pour tenir compte du fait que la distribution est tronquée à zéro). La F_{LV} médiane = 42,50 %. La F_{LV} de référence = 100%. La densité de probabilité de la F_{LV} est superposée aux mêmes données présentées au moyen d'un histogramme. Le « rug plot » en bas montre la distribution des points de données.

5.2.3 Les occlusives labiales-vélaires et l'expressivité : estimation de la fréquence des occlusives labiales-vélaires dans le vocabulaire de base.

Les descriptions de langues individuelles ou de groupes linguistiques mentionnent parfois que les occlusives LV sont plus fréquentes dans les parties expressives du vocabulaire que dans le vocabulaire général. Par exemple, Bostoen & Donzo (2013) soulignent que les occlusives LV en ngombe (bantu, RDC ; [ngom1268]) sont beaucoup plus fréquentes dans les idéophones, les mots dérivés d'idéophones et les adverbes à symbolisme phonétique que dans le lexique général. Ils en concluent que les occlusives LV sont associées à l'expressivité dans cette langue. De même, Martin (2015) souligne que les occlusives LV sont beaucoup plus fréquentes dans les idéophones qu'ailleurs en wawa [wawa1246], une langue mambiloïde parlée au Cameroun, près de la frontière avec le Nigeria. Le ngombe et le wawa diffèrent l'un de l'autre en ce que les occlusives LV sont des phonèmes courants en ngombe, mais rares en wawa. Par conséquent, il peut y avoir une préférence pour les occlusives LV pour se trouver dans les parties expressives du vocabulaire indépendamment de leur fréquence lexicale globale dans une langue.

Un exemple particulièrement parlant d'une telle association préférentielle entre les occlusives LV et l'expressivité est fourni par la variété nkundo du groupe dialectal mongo (bantu, RDC ; [nkun1238], [mong1338]) telle que décrite par Hulstaert (1957; 1961; 1965; 1966). A travers les dialectes du mongo, nous trouvons une variation entre dialectes et parfois au sein d'un même locuteur entre les occlusives LV /k͡p/ et /g͡b/ et les occlusives vélaires labialisées correspondants /kw/ et /gw/. La plupart des lects mongo préfèrent fortement les réaliser comme occlusives LV, mais le nkundo, le lect de référence de la description grammaticale et lexicale de Hulstaert, préfère les réaliser comme occlusives vélaires labialisées (Hulstaert 1957:xiii, 959). Pourtant, il y a un certain nombre de mots en nkundo où seule une réalisation LV est possible (Hulstaert 1957:xiii). Il est remarquable de constater que l'écrasante majorité de ces quelque 40 mots où seulement une réalisation LV est possible dans le dictionnaire de Hulstaert (1957) sont des idéophones, comme *kpótókpòtò* 'idéophone exprimant le son produit par les pantoufles', complété par un couple de noms dérivés d'idéophones, comme *lì-kpòtò* 'pantoufles', un couple de noms émotionnellement chargés pour lesquels aucun idéophone correspondant n'est enregistré, comme *lì-kpéké* 'escroquerie, arnaque' et *ngbàngà* 'querelle, bagarre', et un couple de termes de niveau subordonné, comme *ngbàà* 'type de ciseau avec un long manche pour la sculpture du bois' et *èngbélɛ́* 'type spécial de gâteau de manioc'.

Je cherche à savoir si la fréquence plus élevée des occlusives LV dans les parties expressives du vocabulaire est une tendance générale en Afrique subsaharienne septentrionnale, car cela pourrait nous éclairer sur la manière dont ces consonnes sont transférées d'une langue à l'autre et contribuer ainsi à expliquer la distribution aréale. Comme il n'existe aucun moyen d'extraire automatiquement des listes fiables de vocabulaire expressif à partir de mes sources lexicales, j'ai utilisé une approximation de l'hypothèse en testant si les occlusives LV sont moins fréquentes dans le vocabulaire de base des langues de mon échantillon que dans leur vocabulaire général, sous l'hypothèse généralement admise que les éléments expressifs ne font pas partie du vocabulaire de base. Pour ce faire, j'ai restreint mon échantillon aux 178 langues comportant des occlusives LV pour lesquelles ma source lexicale dans RefLex comporte au moins 400 entrées.[25] À partir de ces sources, j'ai extrait automatiquement des listes de 200 éléments de vocabulaire de base basées sur la liste « Swadesh 200 » (Swadesh 1952). Étant donné que toutes les sources manquaient d'un nombre variable d'entrées dont la traduction correspondait à un élément de la liste Swadesh 200 (de 21 à 139 entrées manquantes, la

[25] Le seuil de 400 entrées est en accord avec les conclusions de Dockum & Bowern (2019:50) qui soutiennent qu'on risque très probablement de représenter de manière incorrecte des faits de base concernant la phonologie de la langue quand on utilise des listes de mots inférieures au seuil de 400 mots.

moyenne étant de 67), j'ai fini avec un échantillon de 178 listes quasi-Swadesh 200 de taille inégale, utilisant au total 196 concepts sur les 200 concepts de la liste Swadesh 200. J'ai ensuite calculé les fréquences lexicales des occlusives LV dans les listes quasi-Swadesh 200 en pourcentage de leur fréquence de référence et je les ai comparées aux fréquences lexicales des occlusives LV dans les sources complètes dont elles ont été extraites.

Les résultats sont présentés dans la figure 33. Elle représente les densités de probabilité des fréquences de F_{LV} en pourcentage dans le sous-échantillon de 178 sources RefLex ayant au moins 400 entrées (en rose) et dans les listes quasi-Swadesh 200 dérivées de ce sous-échantillon (en bleu). La F_{LV} médiane de ces dernières (ligne bleue pointillée) est à 24,95 %, bien inférieure à la F_{LV} médiane de 42,49 % des premières (ligne rouge pointillée), qui est elle-même largement inférieure à la fréquence de référence de 100 % (ligne noire pointillée). Les deux distributions ne sont pas normales, mais leurs variances sont similaires, ce qui permet de les comparer à l'aide du test des rangs signés de Wilcoxon (test U apparié). Ce test confirme qu'il est très peu probable que les deux distributions représentent la même population ($p < .001$) et que la différence entre les valeurs médianes de F_{LV} de ces deux ensembles de données est significative. J'ai aussi effectué une validation par bootstrap (répétitions = 999), qui a également confirmé ce résultat (100% des valeurs $p < .05$, 87% des valeurs $p < .001$). Les résultats résumés dans la figure 33 confirment mon hypothèse : Les occlusives LV sont moins fréquentes dans le vocabulaire de base que dans le vocabulaire général.

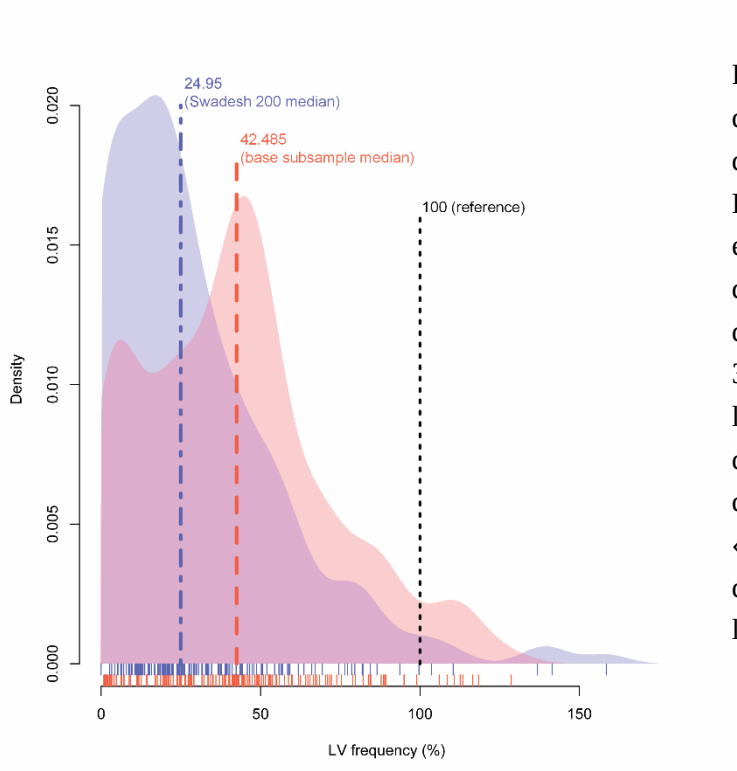


FIGURE 33. Les densités de probabilité des fréquences de F_{LV} en pourcentage dans le sous-échantillon de 178 sources RefLex comportant au moins 400 entrées et dans les listes quasi-Swadesh 200 dérivées de ce sous-échantillon (estimées de la même manière que dans la figure 32). La F_{LV} médiane est de 42,49% dans le sous-échantillon de base et de 24,95% dans les listes quasi-Swadesh 200 dérivées. La F_{LV} de référence = 100%. Les « rug plots » en bas montrent la distribution des points de données dans les deux ensembles de données.

Il est important de noter que les données des listes quasi-Swadesh 200 peuvent également être considérées comme un soutien à mon hypothèse selon laquelle les occlusives LV sont plus fréquentes dans les parties expressives du vocabulaire que dans le vocabulaire général. Ce soutien provient de la distribution des mots avec et sans occlusives LV dans 178 listes quasi-Swadesh 200 par catégorie de partie de discours des concepts, comme résumé dans le tableau 2.[26] La proportion totale de mots avec occlusives LV dans 178 listes quasi-Swadesh 200 est de 4,81% et nous trouvons un rapport similaire de 4,75% pour les noms et les verbes pris ensemble. En même temps, la fréquence des mots avec occlusives LV est presque le double, 8,52%, pour les qualificatifs et les quantificatifs pris ensemble, les deux catégories qui sont les plus susceptibles de contenir des concepts expressifs et évaluatifs. De façon remarquable, nous trouvons la situation inverse avec les catégories de parties du discours qui sont habituellement décrites comme des catégories fonctionnelles, à savoir les conjonctions, les adpositions, les démonstratifs, les pronoms personnels, les marques de négation et les indéfinis. La fréquence des mots avec occlusives LV dans ces catégories fonctionnelles prises ensemble est aussi faible que 0,48%. Les numéraux, qui sont plus une catégorie ouverte semblable aux noms et aux verbes, et les interrogatifs, qui sont semblables aux catégories fonctionnelles mais qui sont canoniquement associés à un statut structurel d'information proéminent, ont des fréquences un peu plus élevées de mots avec des occlusives LV, 1,05% et 2,92% respectivement, que les catégories fonctionnelles plus canoniques prises ensemble.

[26] Puisque, à proprement parler, les catégories de parties du discours sont nécessairement des classes de distribution de lexèmes spécifiques à une langue, la classification des parties du discours utilisée dans le tableau 2 est essentiellement une classification sémantique des concepts enrichie de quelques caractéristiques morphosyntaxiques assez larges. L'étiquette de partie du discours d'un concept donné ne doit pas nécessairement correspondre à l'étiquette de partie du discours utilisée dans la description d'une langue donnée pour le mot traduisant ce concept.

Partie de discours	Concept	Mots avec LV	Mots sans LV	Mots avec LV en %	
conjonction	and, because, if	0	185	0%	
adposition	at, in, with	1	147	0.68%	
demonstratif	here, that, there, this	3	360	0.83%	
pronom	1PL, 1SG, 2PL, 2SG, 3PL, 3SG	1	437	0.23%	**0.48%**
négation	not (NEG)	0	52	0%	
indefini	other, some	1	63	1.56%	
interrogatif	how?, what?, when?, where?, who?	5	472	1.05%	
numéral	five, four, one, three, two	24	798	2.92%	
nom	animal, ashes, back, bark, belly, bird, blood, bone, child, cloud, day, dog, dust, ear, earth, egg, eye, fat, father, feather, fire, fish, flower, fog, foot, fruit, grass, guts, hair, hand, head, heart, husband, lake, leaf, leg, liver, louse, man, meat, mother, mountain, mouth, name, neck, night, nose, person, river, road, root, rope, salt, sand, sea, seed, skin, sky, smoke, snake, star, stick, stone, sun, tail, tongue, tooth, tree, water, wife, wind, wing, woman, woods, worm, year	573	10839	5.02%	
verbe	to bite, to blow, to breathe, to burn, to come, to cut, to die, to dig, to drink, to eat, to fall, to fear, to fight, to float, to flow, to fly, to give, to hear, to hit, to hold, to hunt, to kill, to know, to laugh, to lie, to live, to play, to pull, to push, to rain, to rub, to say, to scratch, to see, to sew, to sing, to sit, to sleep, to smell, to spit, to split, to squeeze, to stab, to stand, to suck, to swell, to swim, to think, to throw, to tie, to turn, to vomit, to walk, to wash, to wipe	300	6684	4.30%	**4.75%**
qualifiant	bad, big, black, cold, dirty, dry, dull, far, good, green, heavy, left, long, narrow, near, new, old, red, right, right (correct), rotten, sharp, short, small, smooth, straight, thick, thin, warm, wet, white, wide, yellow	188	2190	7.91%	
quantifiant	all, few, many	43	290	12.91%	**8.52%**
	TOTAL:	1139	22517	**4.81%**	

Tableau 2. La distribution des mots avec et sans occlusives LV dans 178 listes quasi-Swadesh 200 par la catégorie de partie de discours des concepts.

5.3 L’analyse spatiale

La carte de la figure 30 est une représentation à l’aide d’un motif de points de la distribution géographique des langues africaines comportant des occlusives LV. Elle montre les zones où beaucoup, peu ou pas de langues ont des occlusives LV, ajoutant des détails à une image générale qui était déjà largement connue. Je suis ici principalement intéressé par la recherche de profils dans la distribution géographique des différences dans la fréquence lexicale des occlusives LV parmi les langues qui en possèdent. J’ai donc couplé les résultats de mes estimations de fréquence présentées dans §5.2.2 avec les coordonnées géographiques des langues de mon échantillon. Je commencerai par visualiser les résultats à l’aide de deux types de graphiques d’interpolation spatiale dans §5.3.1, puis je tenterai de quantifier mes résultats de manière plus rigoureuse en les modélisant et en les visualisant à l’aide de la modélisation additive généralisée dans §5.3.2. Enfin, dans §5.3.3, je reproduirai mes résultats en examinant la distribution géographique des toponymes africains qui contiennent des occlusives LV. Les visualisations et les modèles présentés dans la suite de ce chapitre se concentrent sur la zone pour laquelle je dispose de données sur la fréquence lexicale des occlusives LV (intervalle de longitude [-18°, 36°], intervalle de latitude [-9°, 16°]).

5.3.1 L’interpolation spatiale

La figure 34 montre un graphique d’interpolation spatiale des fréquences F_{LV} en pourcentage (y compris 0 % pour les langues sans occlusives LV) produit au moyen d’un lissage gaussien à noyau, tandis que la figure 35 montre un graphique d’interpolation spatiale des fréquences F_{LV} en pourcentage au moyen d’une pondération inverse à la distance. Le lissage par noyau étant un interpolateur inexact, les fréquences F_{LV} lissées de la figure 34 n’atteignent que 100%, contrairement au maximum réel de 140% reflété fidèlement par l’interpolateur exact de la pondération inverse à la distance de la figure 35.

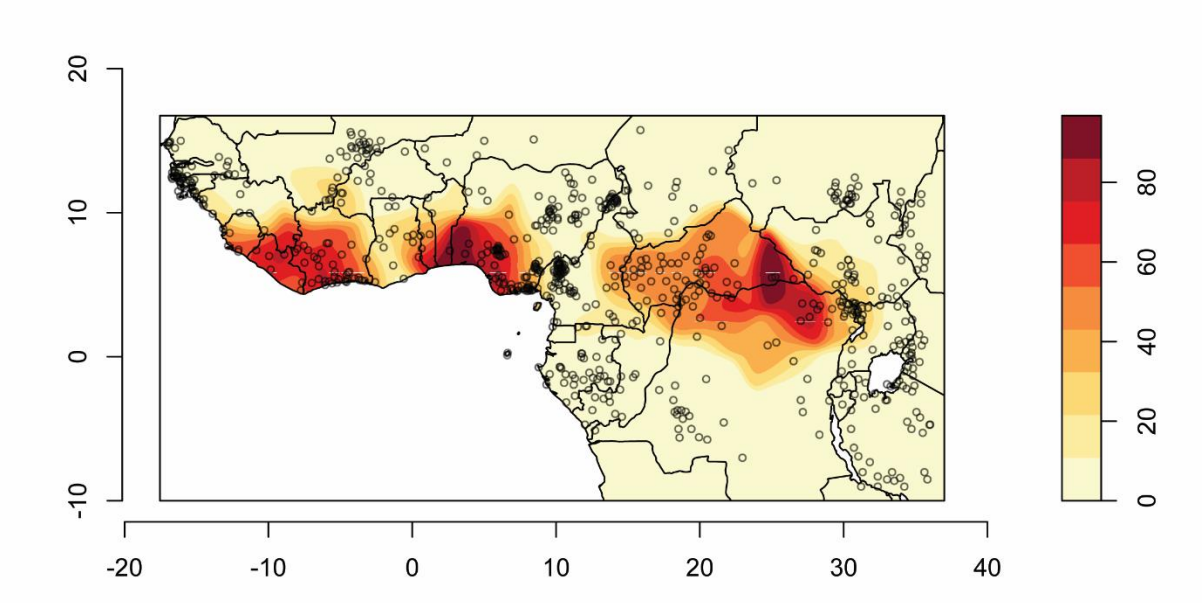


FIGURE 34. Un graphique d'interpolation spatiale des fréquences F_{LV} en pourcentage (y compris 0 % pour les langues sans occlusives LV) produit au moyen d'un lissage à noyau gaussien avec le lisseur de Nadaraya-Watson. La largeur de bande du noyau a été optimisée pour maximiser le critère de validation par recoupement de la vraisemblance du processus ponctuel. Les langues de l'échantillon sont indiquées par des cercles noirs. Le ruban à droite du graphique montre le schéma de couleurs utilisé pour représenter les fréquences F_{LV} lissées en pourcentage.

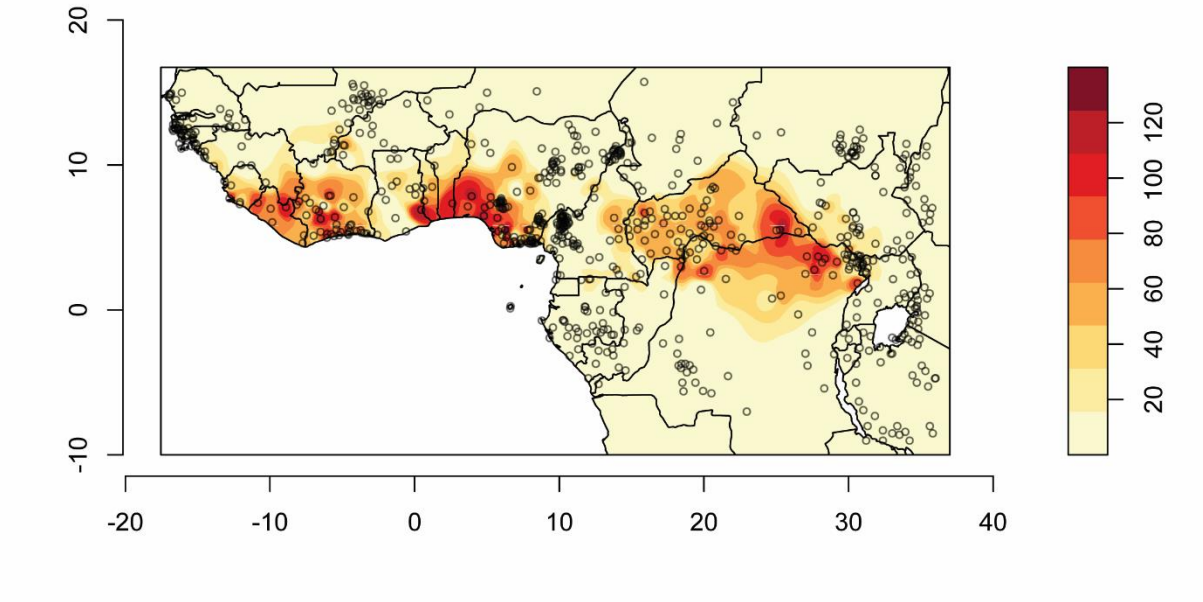


FIGURE 35. Un graphique d'interpolation spatiale des fréquences F_{LV} en pourcentage (incluant 0% pour les langues sans occlusives LV) produit au moyen d'une pondération inverse à la distance (puissance = 5). Les langues de l'échantillon sont indiquées par des cercles noirs. Le ruban à droite du graphique montre le schéma de couleurs utilisé pour représenter les fréquences F_{LV} en pourcentage.

Malgré ces différences mineures, les deux visualisations montrent la même structure générale dans la distribution spatiale des fréquences F_{LV}, le lissage par noyau

de la figure 34 étant naturellement plus adapté pour visualiser les tendances générales. Ainsi, les deux visualisations distinguent clairement trois grandes régions de fréquence lexicale élevée des occlusives LV : (i) la côte du golfe de Guinée à l'ouest du Ghana, que j'appellerai le foyer *Upper Guinea Hotbed*, (ii) la côte du golfe de Guinée à l'est du Ghana, que j'appellerai le foyer *Lower Guinea Hotbed*, et (iii) une zone centrée sur le bassin de l'Oubangui et couvrant la République centrafricaine et le nord de la RDC, que j'appellerai le foyer *Ubangi Basin Hotbed*. Le foyer Lower Guinea Hotbed et le foyer Ubangi Basin Hotbed sont séparés par une grande discontinuité située au Cameroun et au nord-est du Nigeria, que j'appellerai le *Cameroon Gap*. Les deux foyers le long de la côte du golfe de Guinée sont séparés l'un de l'autre par une discontinuité moins importante située dans et autour du Ghana, que j'appellerai le *Ghana Gap*. Puisque dans sa partie sud cette discontinuité correspond principalement à la zone de diffusion de l'akan [akan1250], une grande langue sans occlusives LV, on aurait pu supposer que cette discontinuité n'est qu'apparente et due à la présence accidentelle d'une grande langue sans occlusives LV. Cependant, les deux visualisations montrent clairement l'existence d'un bon nombre de langues à faible fréquence lexicale des occlusives LV au nord de l'aire de diffusion de l'akan, ce qui contribue à l'émergence de cette discontinuité. De plus, l'interpolation par lissage de noyau de la figure 34 suggère l'existence de deux extensions des foyers dont les fréquences de LV sont clairement inférieures à celles des foyers dont elles émergent. La première extension, que j'appellerai la *Banfora Extension*, s'étend du foyer Upper Guinea Hotbed au sud-est du Mali et au sud-ouest du Burkina Faso, en suivant grossièrement l'escarpement de Banfora. La seconde extension, moins importante, que j'appellerai la *Dja-Ntem Extension*, s'étend du foyer Ubangi Basin Hotbed dans les basses terres du sud du Cameroun vers la Guinée équatoriale.

Ce qui est intéressant dans ces visualisations d'interpolation spatiale, c'est qu'ils montrent que l'aire Macro-Sudan Belt / Sudanic zone n'est pas une zone homogène en ce qui concerne la distribution des fréquences lexicales des occlusives LV. §5.5 fournit une interprétation historique de sa structure interne complexe, mais je vais d'abord essayer de quantifier mes résultats d'une manière plus rigoureuse (§5.3.2) et ensuite les valider par recoupement avec un autre ensemble de données (§5.3.3).

5.3.2 La modélisation additive généralisée

5.3.2.1 Les visualisations des modèles additifs généralisés

La figure 36 est une visualisation de la surface de régression GAM des fréquences lexicales des occlusives labiales-vélaires F_{LV} en pourcentage (y compris 0% pour les langues sans occlusives labiales-vélaires).

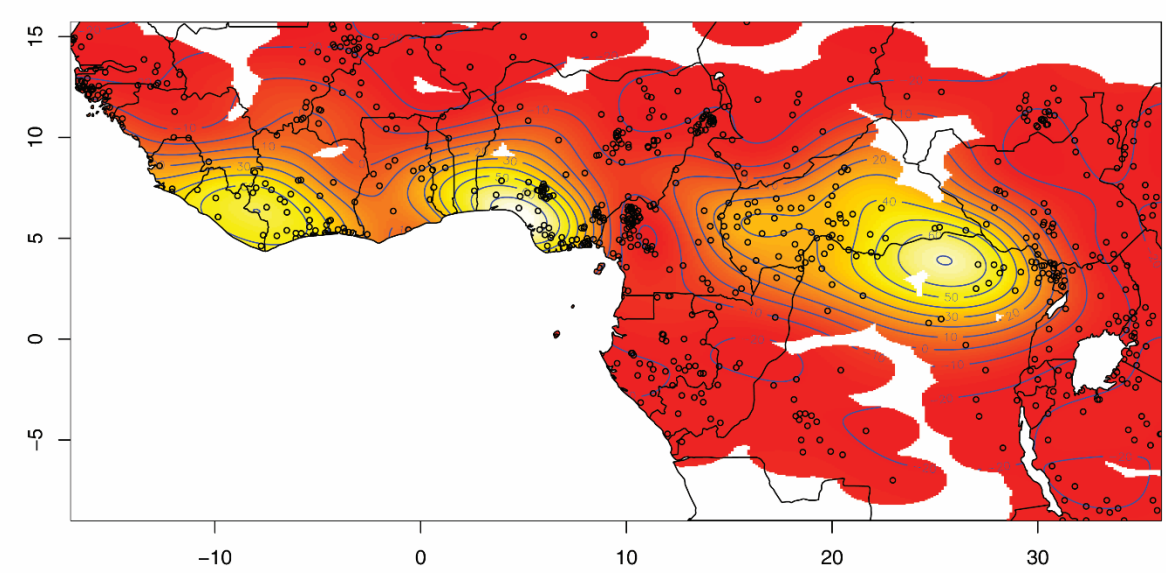


FIGURE 36. Le tracé de contours avec le schéma de couleurs de la carte thermique représentant la surface de régression GAM des fréquences lexicales des occlusives labiales-vélaires (F_{LV}) en pourcentage (y compris 0% pour les langues sans occlusives LV) en fonction de la combinaison de la longitude et de la latitude en utilisant des splines de régression à plaque mince. Le résumé du modèle : k = 13 (k-index = 0.99, p = 0.39, k′ = 195), fonction = gaussienne, edf = 70.76, la déviance expliquée = 77.60%, AIC = 6048, intercept F_{LV} = 19.95%, p < .001.

La visualisation dans la figure 36 est largement comparable aux graphiques d'interpolation spatiale de la figure 34 et de la figure 35 et se prête à des observations similaires à celles qui ont déjà été formulées dans §5.3.1 sur la distribution spatiale des fréquences F_{LV} dans les langues de l'échantillon. Par rapport aux graphiques d'interpolation spatiale, la visualisation GAM confirme davantage la présence du Ghana Gap et met particulièrement en évidence la structure interne des trois foyers. Elle met également en évidence la proéminence particulière des valeurs élevées de F_{LV} dans le Upper Guinea Hotbed par rapport aux deux autres foyers. Les extensions Banfora Extension et Dja-Ntem Extension sont pratiquement absentes de la figure 36.

La robustesse qualitative de la visualisation GAM dans la figure 36 est renforcée par le fait que même les GAMs basés sur des sous-ensembles des données beaucoup plus petits produisent des visualisations très similaires. Ainsi, comparez la visualisation basée sur l'ensemble des données dans la figure 36 avec la figure 37 qui visualise la surface de régression GAM des fréquences F_{LV} en pourcentage uniquement pour les 315 langues pour lesquelles je dispose de données de fréquence lexicale. Bien que la figure 37 soit inévitablement moins précise que la figure 36, elle montre les trois mêmes foyers séparés par les deux mêmes discontinuités.

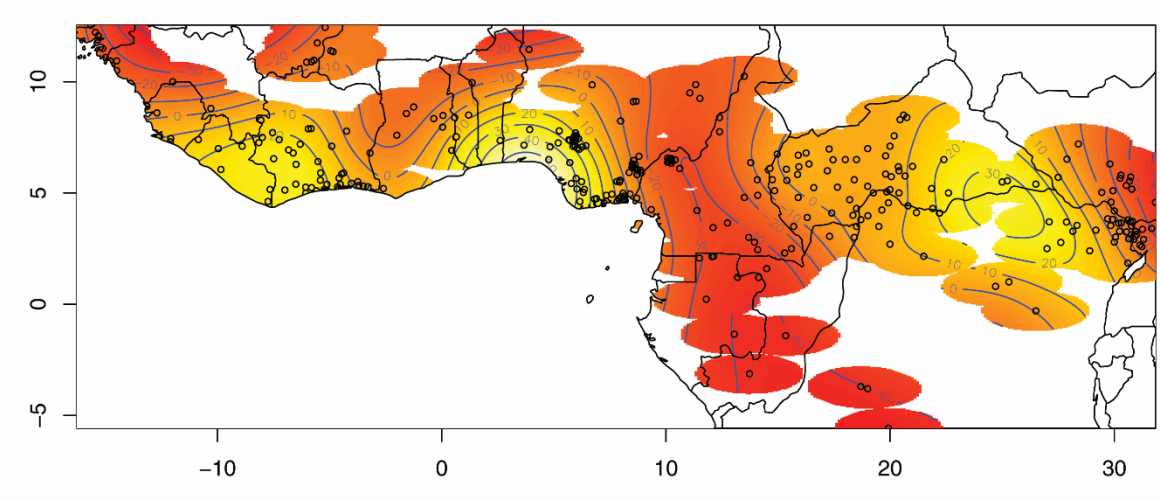


FIGURE 37. Le tracé de contours avec le schéma de couleurs de la carte thermique représentant la surface de régression GAM des fréquences F_{LV} en pourcentage (à l'exception des langues sans occlusives LV avec F_{LV} de 0%) en fonction de la combinaison de la longitude et de la latitude en utilisant des splines de régression à plaque mince. Le résumé du modèle : k = 15 (k-index = 0.99, p = 0.41, k′ = 224), fonction = gaussienne, edf = 29.83, la déviance expliquée = 56.40%, AIC =2831, intercept F_{LV} = 45.92%, p < .001

La figure 38 représente la surface de régression GAM des fréquences F_{LV} en pourcentage pour le sous-ensemble de données qui comprend seulement les 178 langues avec occlusives LV pour lesquelles ma source lexicale dans RefLex a au moins 400 entrées (cf. §5.2.3) plus les langues sans occlusives LV avec F_{LV} de 0%. La figure 38 montre les trois mêmes foyers séparés par les deux mêmes discontinuités, bien que le Ubangi Basin Hotbed ait une forme et une structure plus simples.

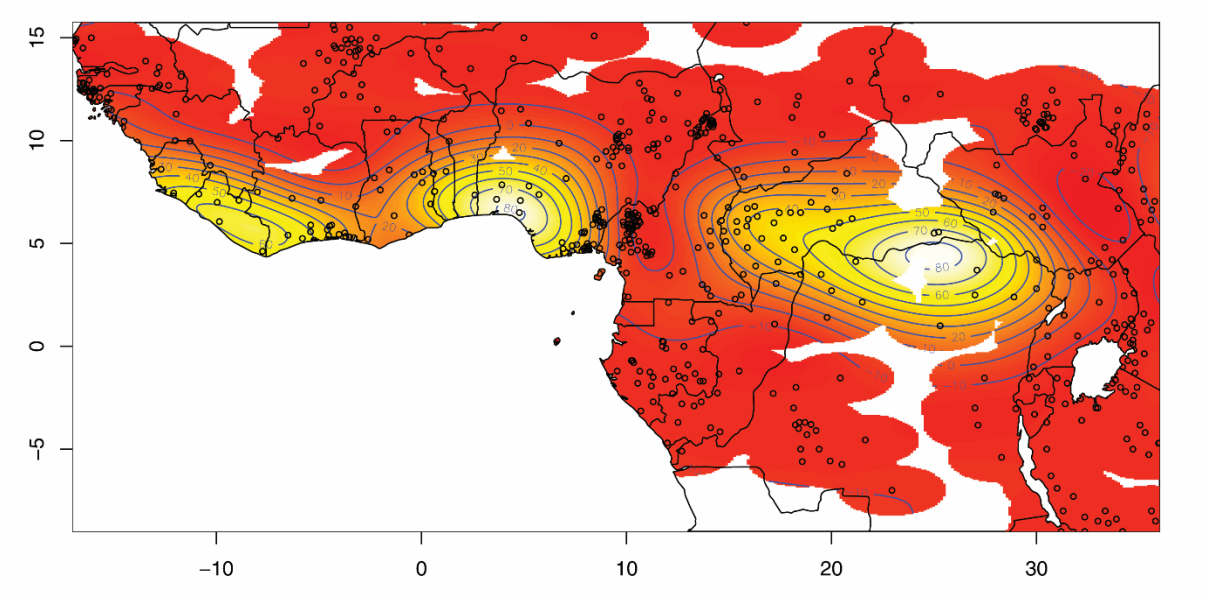


FIGURE 38. Le tracé de contours avec le schéma de couleurs de la carte thermique représentant la surface de régression GAM des fréquences F_{LV} (en pourcentage, en fonction de la combinaison de la longitude et de la latitude en utilisant des splines de régression à plaque mince) pour le sous-ensemble de données qui comprend seulement les 178 langues avec occlusives LV pour lesquelles ma source lexicale dans RefLex a au moins 400 entrées plus les langues sans occlusives LV avec F_{LV} de 0%. Le résumé du modèle : k = 11 (k-index = 1.03, p = 0.76, k′ = 120), fonction = gaussienne, edf = 64.54, la déviance expliquée = 76.80%, AIC = 4760, intercept F_{LV} = 13.40%, p < .001.

La figure 39 représente la surface de régression GAM des fréquences F_{LV} en pourcentage pour le même sous-ensemble que dans la figure 38, mais les valeurs des fréquences F_{LV} sont celles des listes quasi-Swadesh 200 (cf. §5.2.3). La structure spatiale de la figure 39 est très similaire à celle de la figure 38, mais conformément aux effets de la concentration sur le vocabulaire de base discutés dans §5.2.3, les trois foyers deviennent moins importants en termes de valeurs de fréquences F_{LV}. En outre, l'étendue du foyer Ubangi Basin Hotbed diminue légèrement.

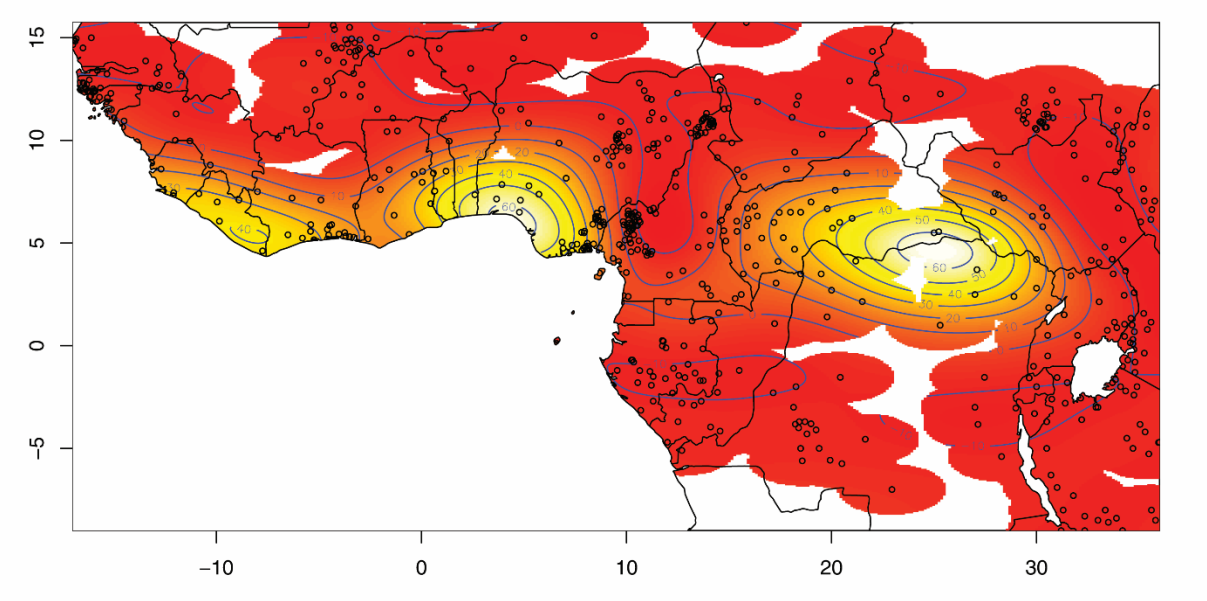


FIGURE 39. Le tracé de contours avec le schéma de couleurs de la carte thermique représentant la surface de régression GAM des fréquences F_{LV} dans les listes quasi-Swadesh 200 (en pourcentage, en fonction de la combinaison de la longitude et de la latitude en utilisant des splines de régression à plaque mince) pour le sous-ensemble de données qui comprend seulement les 178 langues avec occlusives LV pour lesquelles ma source lexicale dans RefLex a au moins 400 entrées plus les langues sans occlusives LV avec F_{LV} de 0%. Le résumé du modèle : k = 11 (k-index = 1.05, p = 0.78, k′ = 120), fonction = gaussienne, edf = 54.49, la déviance expliquée = 61.30%, AIC = 4823, intercept F_{LV} = 9.74%, p < .001.

Enfin, considérons la figure 40 qui est une visualisation de la surface de régression GAM des fréquences F_{LV} log-transformées, y compris les langues sans occlusives LV.[27] L'effet le plus important de la transformation logarithmique pour mes données de fréquence F_{LV} est qu'elle transforme la distribution originale à queue droite (voir la figure 32) en une distribution à queue gauche qui étale les valeurs inférieures des fréquences F_{LV} tout en condensant les valeurs supérieures des fréquences F_{LV}. En raison de cet effet de zoom sur les valeurs inférieures des fréquences F_{LV}, le GAM basé sur les fréquences F_{LV} log-transformées visualise beaucoup mieux les transitions entre les foyers et les zones sans occlusives LV. En même temps, il nivelle la structure interne des foyers et les différences de proéminence entre les foyers, qui sont toutes deux mieux visualisées avec les fréquences F_{LV} en pourcentage non transformées, comme dans la figure 36. La figure 40 met en évidence la nature transitoire du Ghana Gap. Elle confirme également la présence des deux extensions des foyers, la Banfora Extension qui émerge du foyer Upper Guinea Hotbed et la Dja-Ntem Extension qui émerge du foyer Ubangi Basin Hotbed. Finalement, la figure 40 suggère clairement la présence d'un lien entre le foyer Ubangi Basin Hotbed et le foyer Lower Guinea Hotbed le long

[27] Le logarithme de zéro étant indéfini, j'ai mis à l'échelle les valeurs de F_{LV} avant la transformation logarithmique en ajoutant 0,83, la valeur minimale de F_{LV} différente de zéro.

de la vallée du Benue, ce que la figure 36 ne laisse que vaguement entrevoir par la forme des isoplèthes. J'appellerai ce lien entre les deux foyers le lien *Benue River Link*. Une autre observation intéressante concernant le GAM log-transformé visualisé dans la figure 40 est qu'il explique de loin le plus grand pourcentage de déviance (même après avoir pris en compte le nombre légèrement plus élevé de dimensions de base qu'il utilise), à savoir 85,8%. Selon toute probabilité, cette proportion plus élevée de la déviance expliquée est due au fait qu'il y a plus de complexité dans la distribution spatiale des fréquences F_{LV} inférieures par rapport aux fréquences F_{LV} supérieures, qui sont concentrées dans les trois foyers, et qu'en zoomant sur les fréquences F_{LV} inférieures, la log-transformation permet au modèle de tenir compte plus facilement de la complexité de la distribution spatiale de ces valeurs inférieures. Toutes ces propriétés font de la visualisation de la figure 40 la plus informative de manière générale et c'est pourquoi je l'utilise dans le reste du chapitre à titre de référence.

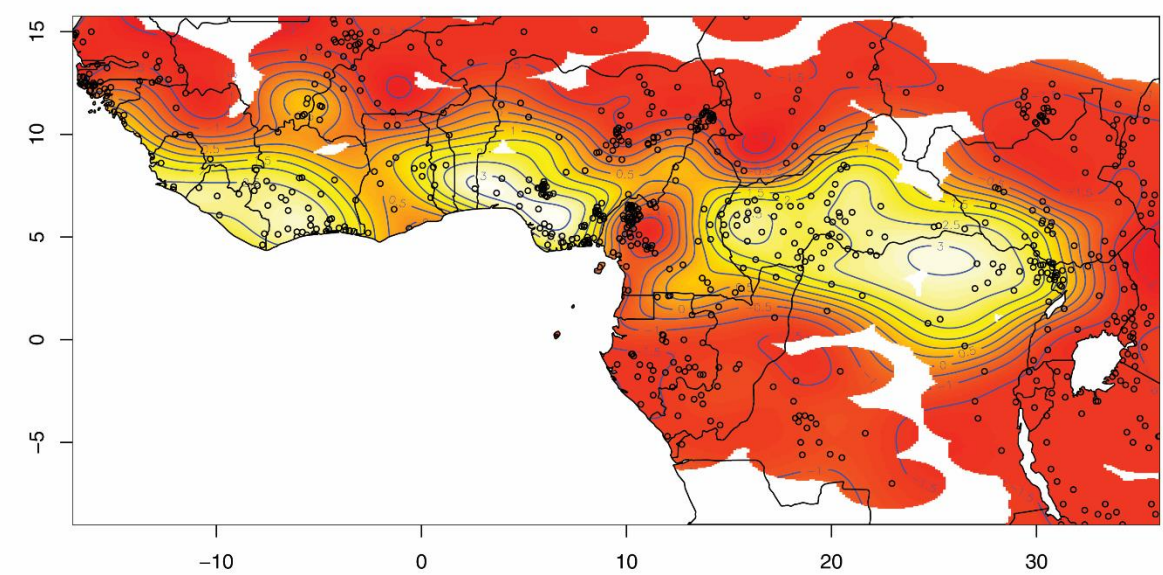


FIGURE 40. Le tracé de contours avec le schéma de couleurs de la carte thermique représentant la surface de régression GAM des fréquences F_{LV} log-transformées (après être mise à l'échelle de 0.83) (y compris les langues sans occlusives LV) en fonction de la combinaison de la longitude et de la latitude en utilisant des splines de régression à plaque mince. Le résumé du modèle : k = 18 (k-index = 1, p = 0.53, k′ = 323), fonction = gaussienne, edf = 108.1, la déviance expliquée = 85.80%, AIC = 1764, intercept log-transformé (après être mise à l'échelle de 0.83) F_{LV} = 1.54837, p < .001.

5.3.2.2 La critique des modèles : le niveau de précision et la robustesse qualitative

Comme j'ai mentioné dans §3.3.2.3, le niveau de précision des estimations des coefficients d'un GAM basé sur la distribution gaussienne dépend de la mesure dans

laquelle le modèle satisfait à la condition que les résidus sont normalement distribués et à la condition que la variance des résidus est constante (homoscédastique) pour toutes les valeurs du prédicteur linéaire. Toutefois, une précision quantitative élevée est beaucoup moins pertinente pour le type de données qu'on examine typiquement dans la typologie aréale, où beaucoup d'imprécision est intrinsèquement intégrée aux données, puisque typiquement la variable dépendante, comme les estimations des fréquences lexicales des occlusives LV, et la variable indépendante, à savoir la combinaison des valeurs de longitude et de latitude prises pour représenter la localisation des langues de l'échantillon, sont nécessairement des approximations grossières. Ce qui importe le plus, c'est la robustesse qualitative des résultats. Dans le cas particulier des résultats présentés ici pour les occlusives LV, cette robustesse qualitative est largement confirmée par une validation croisée utilisant une autre méthode que la modélisation additive généralisé, à savoir l'interpolation spatiale, par une validation croisée avec différents types de sous-échantillons (cf. §5.3.2.1) et par une validation croisée avec un autre type de données, à savoir les toponymes (cf. §5.3.3).

Les estimations des coefficients de la plupart des GAMs présentées dans §5.3.2.2 ne sont pas précises, car deux propriétés de mes données entraînent une violation des hypothèses de normalité et d'homoscédasticité des résidus. La première propriété pertinente est la présence d'un grand nombre de points de données avec des valeurs nulles (langues sans occlusives LV). Ainsi, comme l'illustre la figure 41, qui montre les quatre tracés des résidus pour le GAM de l'ensemble complet de données en pourcentages visualisés dans la figure 36 dans §5.3.2.1, ces points de données nuls forment la ligne oblique lourde dans le tracé Résidus vs Prédicteur linéaire et la ligne horizontale lourde à 0 sur l'axe Réponse dans le tracé Réponse vs Valeurs ajustées, et

ils perturbent fortement la normalité de la distribution dans le tracé Histogramme des résidus.[28]

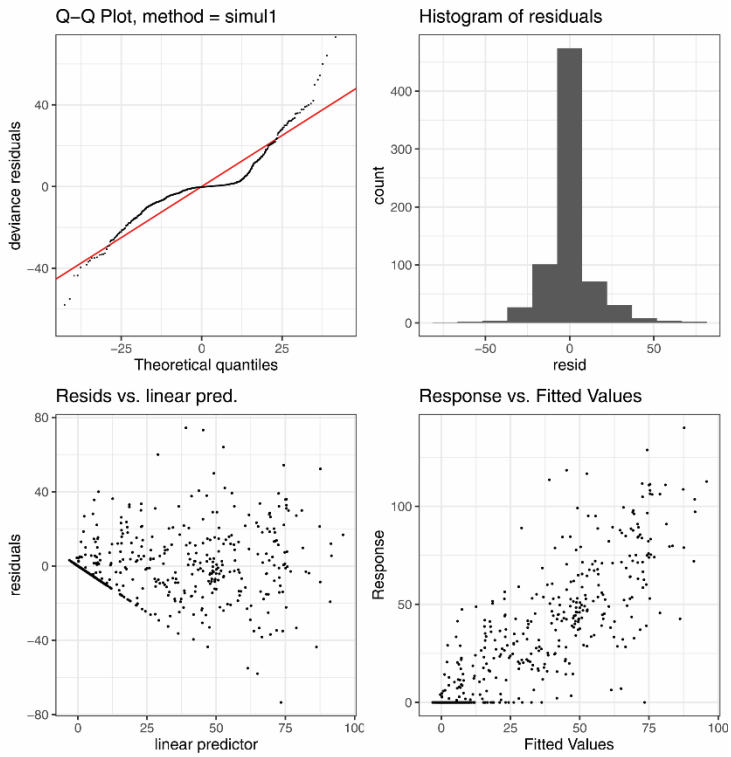


FIGURE 41. Les quatre tracés des résidus pour le GAM de l'ensemble complet de données en pourcentages visualisés dans la figure 36 (§5.3.2.1).

Il n'est donc pas surprenant que la suppression de tous les points de données avec des valeurs nulles (langues sans occlusives LV) de l'ensemble des données comme dans le GAM visualisé dans la figure 37 dans §5.3.2.1 améliore considérablement la distribution des résidus, comme illustré dans la figure 42.

[28] Les ensembles combinés de quatre graphiques de résidus pour les GAMs ont été produits avec le paquet *mgcViz* pour *R* (Fasiolo et al. 2018).

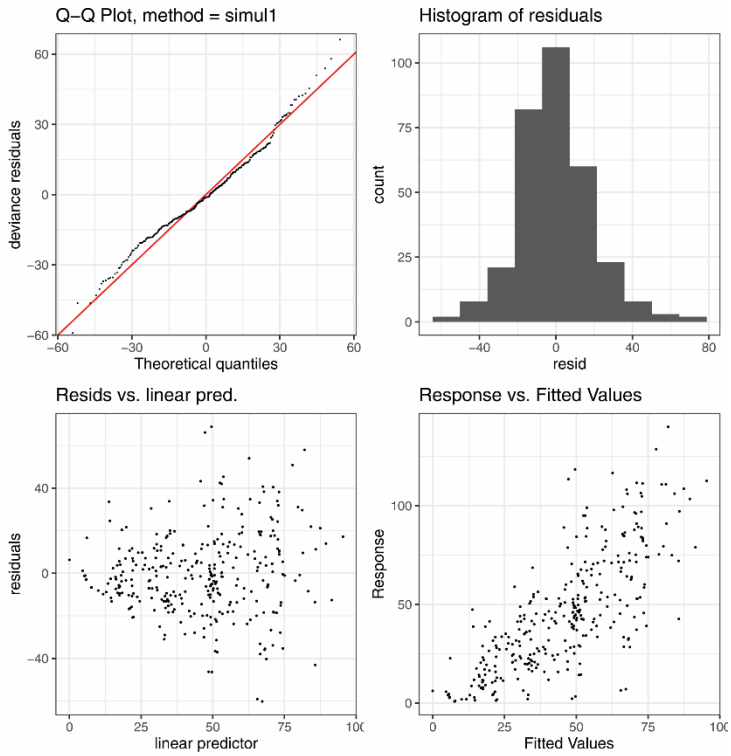


FIGURE 42. Les quatre tracés de résidus pour le GAM de l'ensemble de données en pourcentages avec toutes les langues sans occlusives LV retirées, comme visualisé dans la figure 37 (§5.3.2.1).

En fait, il suffirait maintenant de retirer six résidus aberrants, à savoir les quatre résidus positifs les plus extrêmes et les deux résidus négatifs les plus extrêmes sur le graphique des résidus par rapport au prédicteur linéaire, pour obtenir la normalité dans la distribution des résidus (test de normalité de Shapiro-Wilk : W = 0,99168, p = 0,07979).

Les valeurs aberrantes des résidus mettent en évidence la deuxième propriété de ces données qui entraîne une violation des hypothèses de normalité et d'homoscédasticité des résidus, en particulier dans les modèles qui incluent des langues sans occlusives LV. Cette propriété est la présence de cas de fluctuations locales abruptes des fréquences F_{LV} que j'aborde dans la section suivante §5.3.2.3.

5.3.2.3 Les résidus aberrants : Des fluctuations locales abruptes des fréquences F_{LV} et leur interprétation

Comme j'ai évoqué dans §3.3.2.3, la modélisation additive généralisée peut avoir des problèmes avec de grands changements abrupts dans la valeur de la variable dépendante alors que la valeur de la variable indépendante, c'est-à-dire la combinaison de la longitude et de la latitude, ne change que très peu. Par conséquent, des sauts ou des creux locaux abrupts dans les valeurs de la variable dépendante peuvent entraîner des valeurs aberrantes dans les résidus de la surface de régression produite par un GAM.

La figure 43 est une version du tracé des résidus par rapport au prédicteur linéaire pour le GAM de l'ensemble complet de données en pourcentages (visualisé dans la figure 36 dans §5.3.2.1) présenté dans la figure 41 (§5.3.2.2) qui met en évidence certaines de ces valeurs plus extrêmes des résidus (marquées par des triangles au lieu de cercles). Le tableau 3 associe les indices utilisés dans la figure 43 aux noms des langues respectives.

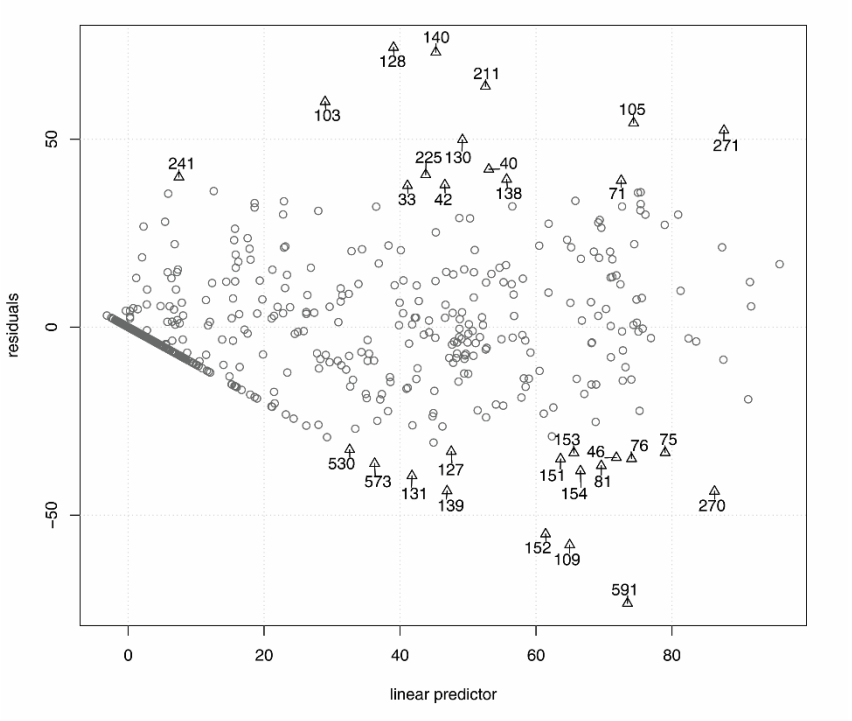


FIGURE 43. Le tracé des résidus par rapport au prédicteur linéaire pour le GAM de l'ensemble des données en pourcentages est visualisé dans la figure 36 (§5.3.2.1) avec certaines des valeurs les plus extrêmes des résidus mises en évidence sous forme de triangles par opposition aux cercles pour les valeurs moins extrêmes des résidus. Les indices sont expliqués dans le tableau 3.

INDEX	LANGUE	GLOTTO-CODE	GENUS	F_{LV} (%)	INDEX	LANGUE	GLOTTO-CODE	GENUS	F_{LV} (%)
33	Monzombo	monz1249	Mundu-Baka	78,75	138	ItuMbuso	itum1245	Cross River	94,99
40	Sherbro	sher1258	Mel	95,06	139	Okobo	okob1241	Cross River	3,34
42	Mbandja	mban1263	Bandaic	84,44	140	Oro	oroo1241	Cross River	118,39
46	Nzakara	nzak1247	Zandic	37,13	151	Ikaan	ukaa1243	Ukaan	28,57
71	Daloa Bete	dalo1238	Kru	111,56	152	Iyinnọ	ukaa1243	Ukaan	6,4
75	Guere	weno1238	Kru	45,63	153	Iigau	ukaa1243	Ukaan	32,15
76	Jrwe	yrew1238	Kru	39,08	154	Ishẹu	ishe1239	Ukaan	28,32
81	Wobe	weno1238	Kru	32,76	211	Logba	logb1245	Kwa	116,62
103	Lendu	lend1245	Lendu	88,94	225	Hai	haii1241	Bandaic	84,34
105	Birri	birr1240	Birri	128,68	241	Baka	baka1272	Mundu-Baka	47,39
109	Gouro	guro1248	Southeastern Mande	7,05	270	Eruwa	eruw1238	Edoid	42,61
127	Ebughu	ebug1241	Cross River	14,46	271	Isoko	isok1239	Edoid	140
128	Efai	efai1241	Cross River	113,44	530	Akan	akan1250	Kwa	0
130	Ekit	ekit1246	Cross River	99,03	573	Ebira	ebir1243	Nupoid	0
131	Enwang	enwa1245	Cross River	2,22	591	Ikwere	ikwe1242	Igboid	0

TABLEAU 3. L'explication des indices des valeurs les plus extrêmes des résidus mis en évidence par les triangles de la figure 43.

En fait, la présence des points de données qui sont susceptibles de poser des difficultés à la modélisation additive généralisée parce qu'ils apportent des fluctuations locales abruptes des fréquences F_{LV} peut déjà, dans une certaine mesure, être déduite du simple graphique d'interpolation spatiale des fréquences F_{LV} en pourcentages au moyen de la pondération inverse à la distance dans la figure 35 (§5.3.1), car cette méthode d'interpolation spatiale permet de mettre en évidence les détails les plus fins de la structure spatiale des données. Pour illustrer ce point et montrer la localisation des langues en question au sein de l'Afrique subsaharienne septentrionale, je reproduis ici la figure 35 en tant que figure 44 et marque par des triangles les valeurs les plus extrêmes des résidus dans le graphique Résidus vs. Prédicteurs linéaires de la figure 43 (voir le tableau 3 pour la signification des indices).

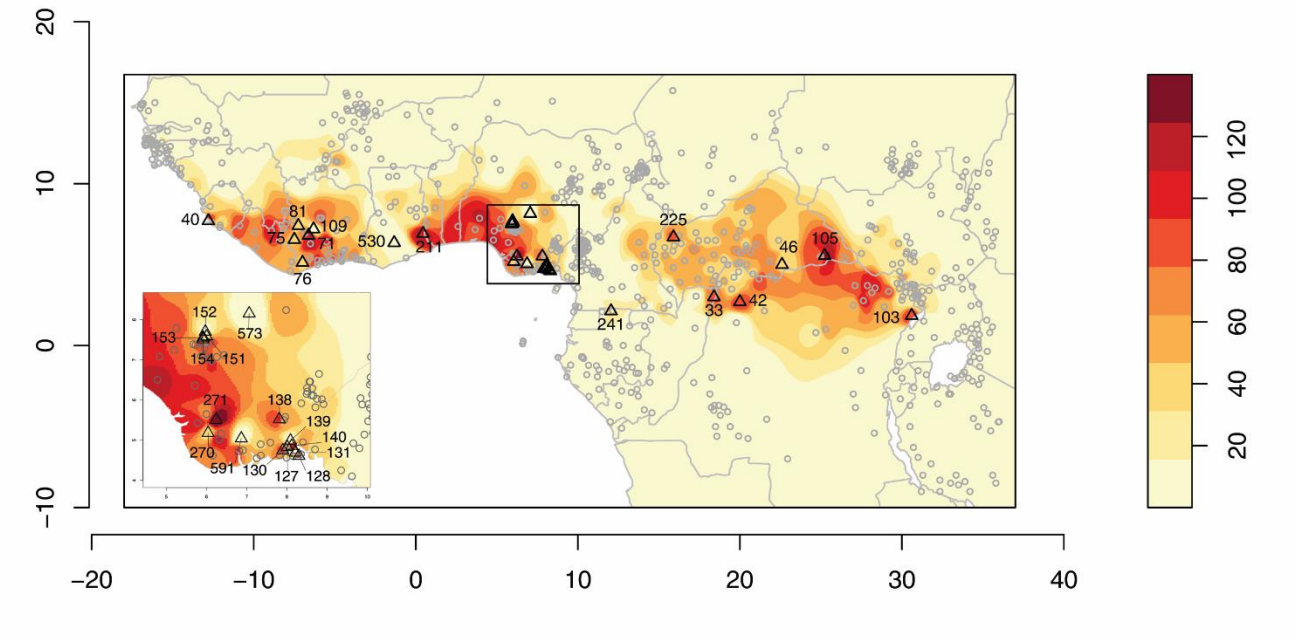


FIGURE 44. Un graphique d'interpolation spatiale des fréquences F_{LV} en pourcentage (incluant 0% pour les langues sans occlusives LV) produit au moyen d'une pondération inverse à la distance (puissance = 5), comparable à la figure 35. Les triangles marquent les valeurs les plus extrêmes des résidus dans le graphique Résidus vs. Prédicteurs linéaires de la figure 43. Voir le tableau 3 pour la signification des indices.

Ces fluctuations locales abruptes de la valeur dépendante, la fréquence F_{LV}, que nous pouvons repérer si facilement en utilisant le graphique des résidus par rapport au prédicteur linéaire d'un GAM permettent à mettre en évidence les sources qui probablement donnent des estimations moins précises de cette valeur dépendante. Les estimations provenant de ces sources doivent être validées par recoupement avec de meilleures sources, si elles sont disponibles. La principale origine possible de ces estimations inadéquates est la taille particulièrement petite de certaines des sources. Lorsqu'une source est petite, même des changements mineurs peuvent avoir un effet significatif sur sa valeur F_{LV}. Ainsi, une augmentation ou une diminution accidentelle du nombre de racines avec occlusives LV d'une ou deux racines est peu susceptible d'avoir un effet sur la valeur F_{LV} d'une source avec de nombreuses entrées, alors que les conséquences peuvent être beaucoup plus importantes dans une source avec un petit nombre d'entrées. Parmi les langues listées dans le tableau 3, cette possibilité doit être envisagée pour le nzakara, dont la source ne compte que 225 entrées, le monzombo, dont la source ne compte que 240 entrées, les deux langues edoïdes, car elles comptent environ 172 et 177 entrées, les quatre langues ukaan, car mes sources ukaan comptent environ 200 mots par source, trois des quatre langues kru, à savoir le guere, le jrwe et le wobe, car les quatre sources kru ont environ 300 mots par source, mais les trois langues en question ont une fréquence F_{LV} nettement inférieur à celui des langues environnantes dans mon échantillon.

En outre, les fluctuations locales abruptes de la valeur dépendante témoignent très probablement des événements historiques relativement peu profonds. Si de telles fluctuations abruptes s'étaient produites à des profondeurs temporelles plus significatives, nous nous attendrions à ce qu'elles aient été davantage lissées à l'heure actuelle. Deux types de processus spatio-temporels principaux peuvent se manifester, à savoir des événements récents de perte ou d'émergence locale du trait et des événements récents de propagation ou de relocalisation de langues qui auraient fait converger dans une petite région des langues présentant des profils de trait en question sensiblement différents (y compris l'absence totale du trait). Ces processus peuvent alors produire deux types de configurations dans la distribution spatiale de la valeur dépendante qui peuvent être conceptualisés comme des pics et des falaises positifs ou négatifs respectivement.

Si le processus de perte partielle ou complète d'occlusives LV n'affecte qu'une seule langue de l'échantillon dans une zone avec des valeurs F_{LV} relativement élevées, il en résultera un pic négatif dans les valeurs F_{LV} à ce point de données. Par exemple, c'est la raison qui explique la chute abrupte de la valeur F_{LV} jusqu'à zéro pour la langue igboïde l'ikwere, entourée de points de données avec des valeurs F_{LV} élevées, ce qui donne lieu à la valeur aberrante (négative) la plus extrême dans mon GAM.[29] Une valeur aberrante moins radicale dans les résultats du GAM est représentée par la langue nupoïde l'ebira. Dans le dialecte de l'ebira de mon échantillon, les occlusives LV ont récemment changé pour des occlusives labiales (cf. Scholz 1976:8),[30] alors que les langues autour de l'ebira dans mon échantillon ont des valeurs F_{LV} relativement élevées. Un pic négatif local dans les valeurs F_{LV} peut également refléter la situation où les occlusives LV se sont développées au sein d'un groupe de langues apparentées parlées dans la même région générale, mais où une langue a pris du retard dans cette évolution. Dans mes données, un exemple d'une telle évolution peut être représenté par l'iyinno, un dialecte de l'ukaan. Il est possible que la même explication s'applique au pic négatif des valeurs de F_{LV} dans la langue édoïde l'eruwa, dont la valeur relativement faible de F_{LV} contraste fortement avec un pic positif local des valeurs F_{LV} dans une langue édoïde voisine, l'isoko. Cependant, comme mentionné ci-dessus dans les cas de l'ukaan et de l'edoïd, ces fluctuations locales abruptes des valeurs F_{LV} peuvent finalement s'avérer être dues à la petite taille des sources respectives. Un pic local dans les valeurs F_{LV} peut également refléter une migration récente d'une langue dans une zone avec un profil F_{LV}

[29] En fait, l'ikwere n'est pas la seule langue igboïde à avoir subi le processus de perte partielle ou complète des occlusives LV (cf. Blench & Williamson 2016:13), mais c'est la seule langue de ce type de mon échantillon dans cette région.

[30] A cet égard, il convient de signaler une orthographe alternative du nom de cette langue, *egbira*, qui reflète la présence d'une occlusive labiale-vélaire dans d'autres dialectes.

différent. C'est l'explication du pic positif local dans les valeurs F_{LV} produites par le hai, une langue langue banda, dans le nord-ouest de la République centrafricaine. La valeur F_{LV} du hai est comparable à celle des autres langues banda de mon échantillon, mais elle est significativement plus élevée que les valeurs F_{LV} des langues de mon échantillon qui entourent le hai et nous savons que les locuteurs du hai ont migré vers leur emplacement actuel depuis une zone située à au moins 500 km à l'est dans la première moitié du 19ème siècle (cf. Moñino 2004:28). Le pic positif des valeurs F_{LV} dans le coin sud-est de la République centrafricaine produit par le birri peut être particulièrement pertinent d'un point de vue historique car le birri est un isolat potentiel. Malheureusement, on sait actuellement très peu de choses sur cette langue et son histoire (cf. Güldemann 2018a:269, 359).

Dans le sud-est du Nigeria, près de la frontière avec le Cameroun (voir l'encart de la figure 44), nous observons un entrecroisement inhabituel de pics positifs et négatifs des valeurs F_{LV} dans une zone très limitée produite par plusieurs langues de la branche Lower Cross du groupe Cross River. Ce profil enchevêtré est le résultat d'un processus de perte des occlusives LV qui a affecté plusieurs de ces langues, telles que l'ebughu, l'enwang et l'okobo à des degrés divers, tandis que des valeurs élevées de F_{LV} ont été préservées dans les autres langues, telles que l'efai, l'ekit, l'ituMbuso et l'oro.[31] Le processus de perte des occlusives LV parmi les langues Lower Cross doit être relativement récent, car il n'a pas encore affecté toutes ces langues dans une mesure similaire.

Les transitions entre les foyers de fortes valeurs F_{LV} et les zones sans occlusives LV sont généralement graduelles, ressemblant à la pente d'une colline. Cependant, dans un certain nombre de cas, nous observons des transitions abruptes, ressemblant davantage à la face d'une falaise. Comme les pics positifs et négatifs des valeurs F_{LV}, ces falaises causent des difficultés pour la modélisation additive généralisée, comme le montre le graphique Résidus vs Prédicteurs linéaires. Comme les pics, les falaises peuvent également refléter une variété d'événements historiques relativement récents. Par exemple, un certain nombre de langues kwa dans mes données, comme l'akan, qui ont partiellement ou complètement perdu les occlusives LV, contribuant ainsi à l'émergence du Ghana Gap, se trouvent à la frontière des langues kwa de la périphérie occidentale du foyer Lower Guinea Hotbed avec des valeurs F_{LV} élevées, comme le logba, avec pour résultat une fluctuation des valeurs F_{LV} semblable à une falaise. Voir §5.5.3 pour une discussion des mouvements de populations qui ont pu contribuer à l'émergence de cette falaise F_{LV}. Un exemple comparable est fourni par la langue mande

[31] Par exemple, le mot 'léopard' a été reconstruit pour le proto-Lower Cross par Connell (1991) comme **é-kpè* parce que ses réflexes ont *kp* dans presque toutes les langues Lower Cross, tandis que l'ebughu a *é-píè*, l'enwang *é-pè* et l'okobo *é-pì*.

sud-est, le guro, qui semble avoir perdu une bonne partie de ses occlusives LV et qui a connu un événement de propagation assez important dans la périphérie nord-est du foyer Upper Guinea Hotbed, tous deux relativement récents. Une autre falaise F_{LV} se trouve à l'extrémité sud-est du foyer Ubangi Basin Hotbed, où le lendu, une langue du groupe lendu, avec une valeur F_{LV} élevée à l'ouest de la chaîne de montagnes à la frontière entre la RDC et l'Ouganda, est voisin de langues bantu sans occlusives LV à l'est et au sud, et de langues moru-madi et nilotiques avec des valeurs F_{LV} faibles ou sans occlusives LV à l'est et au nord. Tous ces voisins sont connus pour s'être déplacés dans la zone située à l'est de la chaîne de montagnes, soit en provenance du sud (les langues bantu), soit en provenance du nord (les langues nilotique et moru-madi).

5.3.3 Une validation des résultats par recoupement : Les occlusives labiales-vélaires dans les toponymes africains

L'analyse spatiale détaillée des fréquences lexicales des occlusives LV dans les langues africaines que j'ai présentée dans les sections précédentes a été rendue possible par l'existence de la base de données RefLex. RefLex a fourni des données lexicales de quantité et de qualité suffisantes pour soixante pour cent des 545 langues de mon échantillon global qui ont des occlusives LV. Malheureusement, cela laisse de côté quarante pour cent des langues concernées, soit parce qu'il n'existe aucune source lexicale appropriée, soit parce qu'aucune source appropriée n'a encore été incluse dans RefLex. De plus, la taille des sources disponibles dans RefLex est très variable, et certaines régions et familles sont relativement mieux représentées que d'autres. C'est pourquoi j'ai jugé utile de procéder à une validation de mes résultats par recoupement avec un autre ensemble de données, à savoir les toponymes africains figurant dans la base de données GeoNames (GeoNames.org). Plus précisément, j'ai étudié la distribution spatiale des toponymes uniques écrits avec des graphèmes susceptibles de représenter une occlusive LV (par exemple *kp*, *gb* ou Yoruba *p*),[32] en supposant que la

[32] Il y a beaucoup de faux doublets de toponymes dans la base de données GeoNames, beaucoup plus que de vrais doublets. Pour éliminer l'effet des faux doublets de toponymes dans mon ensemble de données, j'ai arbitrairement conservé uniquement le premier point de données pour tout ensemble de doublets de toponymes. J'ai également supprimé tous les toponymes écrits avec *gb* ou *kp* en Afrique du Sud et en Namibie, comme *Springbok*, puisque ces toponymes d'origine germanique ne contiennent pas d'occlusives LV. En même temps, je n'ai pas pris la peine de supprimer manuellement les quelques toponymes écrits avec *gb* et *kp* à la périphérie nord de l'Afrique subsaharienne septentrionale qui ne peuvent pas non plus contenir d'occlusives LV puisqu'il n'y a pas de langues avec des occlusives LV dans ces zones (et il n'y en a pas eu non plus depuis le début de la colonisation européenne lorsque les premières cartes de la région ont commencé à être produites).

fréquence des occlusives LV dans cette partie du lexique devrait globalement être en corrélation avec leur fréquence dans le lexique général. Les résultats sont présentés dans la figure 45, que nous pouvons comparer avec la visualisation du GAM dans la figure 40 reproduite ici dans l'encart.

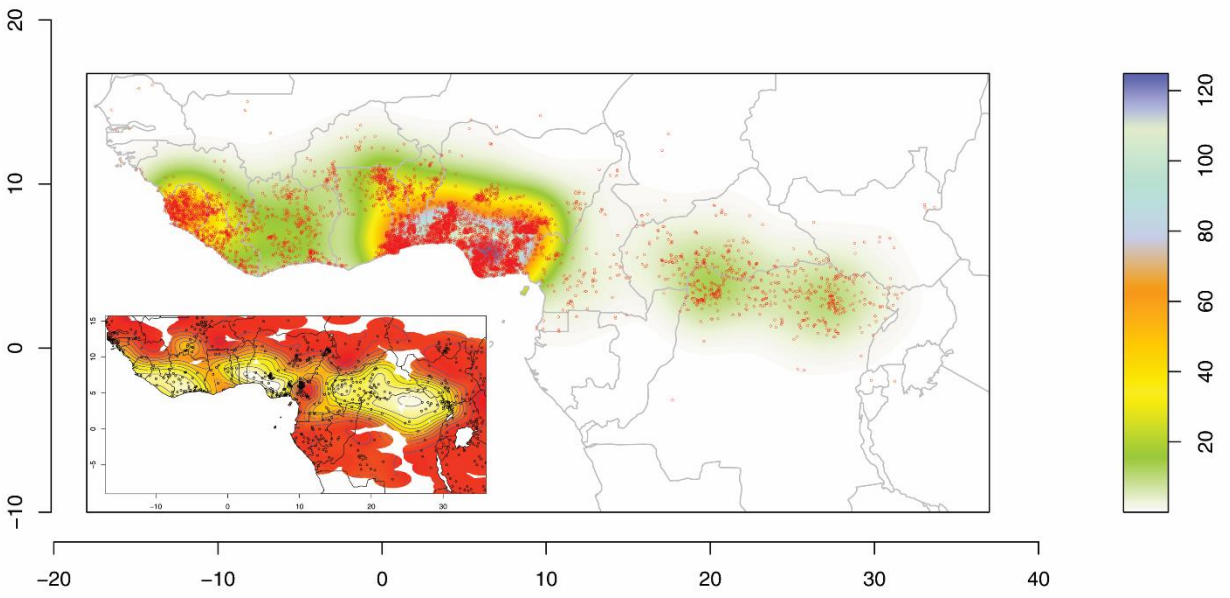


FIGURE 45. La distribution géographique des toponymes uniques écrits avec des graphèmes susceptibles de représenter une occlusive LV dans la base de données GeoNames (GeoNames.org) (cercles rouges) et leur intensité spatiale. Pour comparaison, l'encart reproduit la figure 40 visualisant le GAM des valeurs de F_{LV} log-transformées.

Comme nous pouvons le voir, la distribution spatiale des toponymes écrits avec des graphèmes susceptibles de représenter une occlusive LV dans la figure 45 est caractérisée par la présence de trois régions de haute intensité spatiale de toponymes écrits avec des graphèmes susceptibles de représenter une occlusive LV qui correspondent clairement aux trois foyers de haute fréquence lexicale des occlusives LV dans la visualisation GAM. Comme dans mon jeu de données des fréquences F_{LV}, nous pouvons à nouveau observer un lien plus fort entre le foyer Upper Guinea Hotbed et le foyer Lower Guinea Hotbed et un lien beaucoup plus faible entre le foyer Lower Guinea Hotbed et le foyer Ubangi Basin Hotbed. Une différence majeure est que ni les deux extensions des foyers, ni le lien Benue River Link ne sont visibles dans le jeu de données des toponymes. Cependant, au moins pour le Benue River Link, cette différence n'est qu'apparente et est due à deux raisons. Premièrement, la couverture des données de la base de données GeoNames est déficiente pour les toponymes de la vallée du fleuve

Benue au Nigeria.[33] Deuxièmement, et plus important encore, de grandes étendues de terres le long du fleuve Benue ont été colonisées au cours des derniers siècles par des locuteurs du peul [fula1264], une langue sans occlusives LV parlée à l'origine au Sénégal et dans les régions voisines. L'incursion des Peuls dans la région du fleuve Benue a été à la fois une migration de bergers nomades et une expansion militaire et religieuse dont le résultat a été la création de l'émirat peul d'Adamawa qui contrôlait la majeure partie de la région concernée. Étant donné le statut socio-politique dominant des Peuls dans une grande partie de la région, il est fort probable que de nombreux toponymes employés auparavant dans les zones actuellement habitées par les Peuls ont été remplacés par de nouveaux toponymes peuls, ou du moins adaptés à la prononciation peul. En outre, un nombre important de toponymes dans les zones encore habitées par des locuteurs de langues minoritaires qui ont souvent des occlusives LV apparaissent sur les cartes officielles dans des formes simplifiées sans occlusives LV.[34]

Comme dans le jeu de données des fréquences lexicales F_{LV} (cf. la figure 36 dans §5.3.2.1), dans le jeu de données des toponymes, le foyer Lower Guinea Hotbed est particulièrement proéminent par rapport aux deux autres foyers. Dans le même temps, le foyer Ubangi Basin Hotbed semble beaucoup plus faible dans le jeu des données des toponymes de la figure 45 que dans le jeu de données des fréquences lexicales F_{LV}. Afin d'apprécier la véritable importance du foyer Ubangi Basin Hotbed dans le jeu des données des toponymes, nous devons comparer la distribution géographique des toponymes uniques écrits avec des graphèmes susceptibles de représenter une occlusive LV à la distribution des toponymes écrits sans graphèmes susceptibles de représenter une occlusive LV, comme dans la figure 46. La figure 46 suggère clairement que la faiblesse apparente de ce foyer dans le jeu de données des toponyms de la figure 45 est un artefact de la faible densité de population générale en Afrique centrale, comme le reflète la faible densité des lieux habités (comparer §4.5.2 sur une zone à peu près similaire dans la discussion de la typologie aréale des marques de négation en fin de

[33] Une comparaison rapide, que j'ai effectuée, de quelques zones de la vallée du fleuve Benue dans la base de données GeoNames avec les cartes au 1:100.000 par Nigeria Federal Surveys (1958–1973), comme la feuille 175 Shellen suggère que la couverture de la base de données GeoNames peut parfois descendre jusqu'à 50% des toponymes pour cette partie du Nigeria.

[34] Voir par exemple Shimizu (1979:64) sur les noms de lieux mumuye avec /k͡p/ mal représenté comme *p* sur les cartes.

phrase). La figure 46 confirme de la même manière l'importance des deux discontinuités entre les trois foyers.

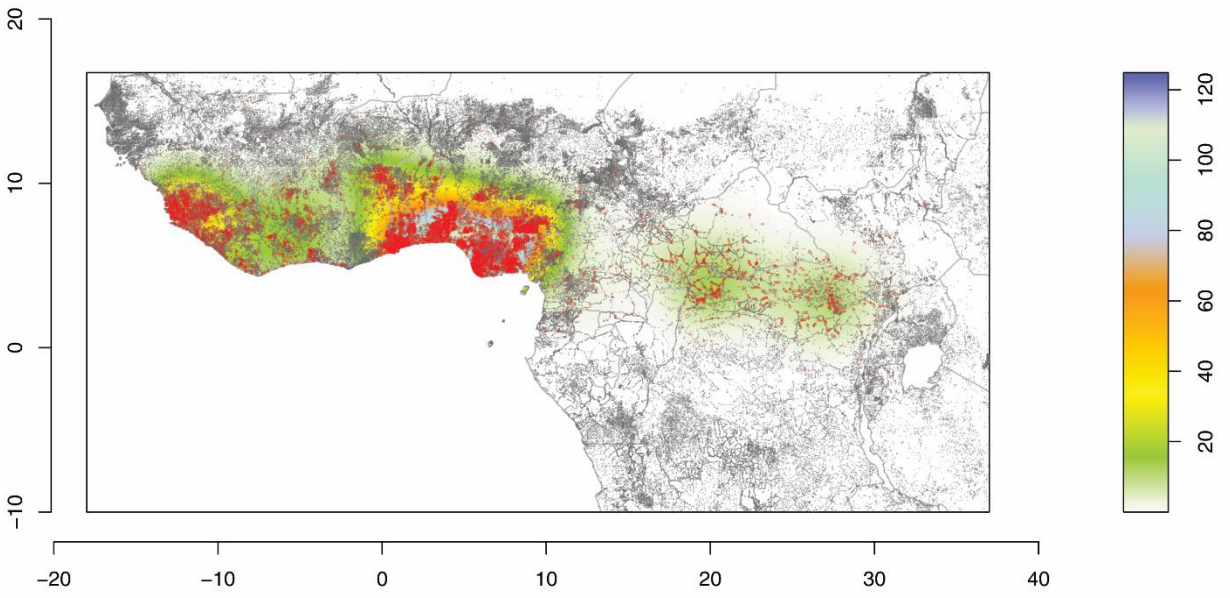


FIGURE 46. La distribution géographique des toponymes uniques écrits avec des graphèmes susceptibles de représenter une occlusive LV (cercles rouges), avec leur intensité spatiale, et ceux écrits sans graphèmes susceptibles de représenter une occlusive LV (cercles gris) dans la base de données GeoNames (GeoNames.org).

5.4 La prosodie d'emphase-C peut expliquer l'émergence, la propagation et la distribution intralinguistique des occlusives labiales-vélaires

Dans §5.2.3, j'ai montré que les occlusives LV sont moins fréquentes dans le vocabulaire de base représenté dans les listes Swadesh 200 que dans le vocabulaire général des langues individuelles et que dans les listes Swadesh 200, les occlusives LV sont particulièrement rares dans les mots fonctionnels, et significativement moins fréquentes dans les concepts nominaux et verbaux par rapport aux concepts qualifiants et quantifiants plus expressifs. J'ai soutenu qu'il s'agit d'une confirmation quantitative indirecte de l'observation selon laquelle les occlusives LV sont relativement plus fréquentes dans les parties expressives du vocabulaire. Un autre type d'asymétrie intralinguistique dans la distribution des occlusives LV qui a souvent été noté dans la littérature est leur forte tendance à se limiter à la position initiale du radical (cf. Connell 1994:468; Cahill 2018:151). Cette tendance est un cas spécifique d'une tendance plus générale d'asymétrie phonotactique dans les langues de l'Afrique subsaharienne septentrionale, qui se manifeste par un nombre décroissant d'oppositions phonologiques

et une application croissante des règles de lénition à mesure que l'on s'éloigne de la position consonantique initiale du radical (C_1), que ce soit vers la droite ou – dans les langues à préfixe – vers la gauche (par exemple, Hyman 2004:80–81; Lionnet & Hyman 2018:652–655).

Les deux types d'asymétrie intralinguistique peuvent être expliqués par ce que j'appelle la prosodie d'emphase-C, un type de proéminence prosodique au niveau du mot ou de l'énoncé dont le corrélat phonétique principal est la longueur de la consonne. Dans un certain nombre de langues bantu du nord-ouest, ce phénomène a été décrit en termes d'accent de mot sur la consonne initiale des racines (C_1), par Paulian (1975) pour le kukuya [teke1280] et par Van de Velde (2008) pour l'eton [eton1253]. La figure 47 illustre un tel accent de mot réalisé par la longueur de la consonne initiale de la racine dans le mot nonsensique *mə-màmà* tel que prononcé par un locuteur eton, où *mə-* est un préfixe de classe et *-màmà* une racine nonsensique.

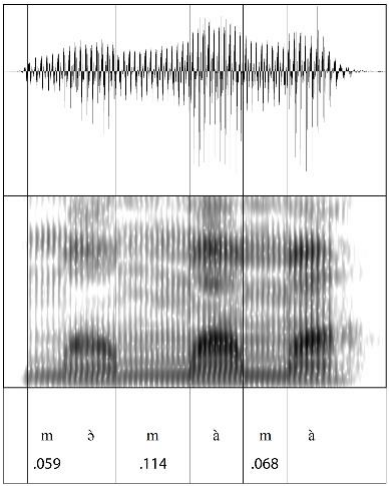


FIGURE 47. Une répétition randomisée par un locuteur de l'eton du mot nonsensique *mə-màmà*, où *mə-* est un préfixe de classe et *-màmà* une racine nonsensique. La durée des trois consonnes *m* est mesurée en secondes.

Les premiers résultats de mes recherche en cours (en collaboration avec Mark Van de Velde) dans un certain nombre de langues d'Afrique de l'Ouest et du Centre suggèrent fortement que la prosodie d'emphase-C est à l'origine un phénomène prosodique/intonationnel au niveau de l'énoncé, marquant une emphase particulière sur un élément donné dans l'énoncé (cf. Idiatov & Van de Velde 2016). Ainsi, la figure 48 illustre la focalisation corrective sur une voyelle de préfixe en eton [eton1253] qui est toutefois réalisée au moyen de l'allongement de la consonne précédente tandis que la voyelle de préfixe elle-même n'est pas affectée.

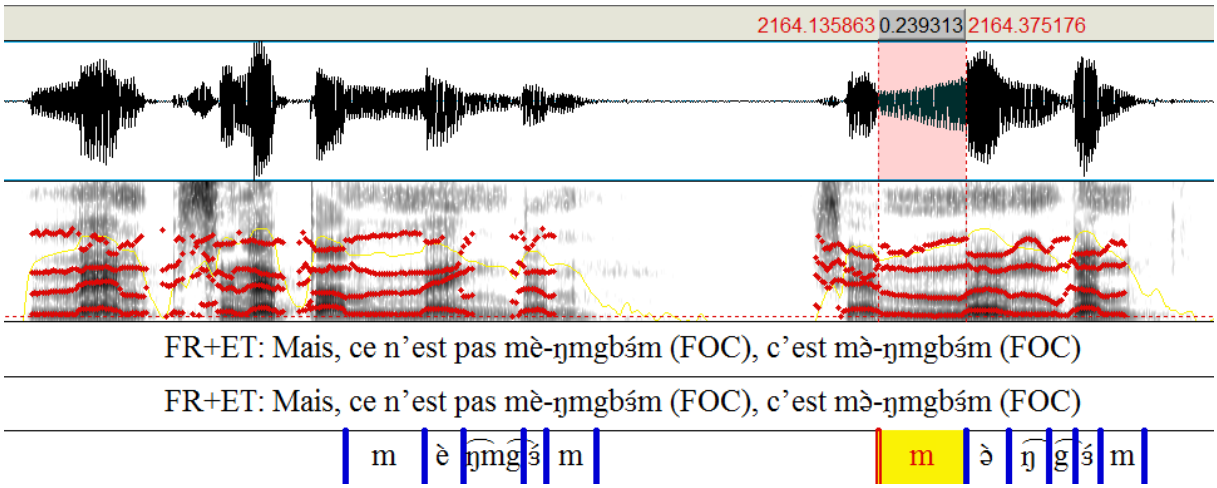


FIGURE 48. Une illustration de la focalisation corrective sur une voyelle de préfixe en eton [eton1253] réalisée au moyen de l’allongement de la consonne précédente tandis que la voyelle de préfixe elle-même n’est pas affectée (Idiatov & Van de Velde 2016).

L’allongement consonantique facilite l’émergence des occlusives LV à partir des vélaires labialisées de deux manières. Premièrement, l’effort articulatoire supplémentaire nécessaire à sa production augmente les chances de dépassement dans la réalisation de l’approximation labiale d’une vélaire labialisée en tant que fermeture complète (comparer Bybee & Easterday 2019:294–295 sur la base articulatoire du renforcement des glides par dépassement). Deuxièmement, la durée supplémentaire augmente les chances que le geste vélaire initial ne finisse pas par être masqué par le geste labial ultérieur. L’utilisation intonative de la prosodie de l’emphase-C comme moyen de marquer l’emphase sur un élément de l’énoncé explique naturellement la prépondérance des occlusives LV dans les parties expressives du vocabulaire. De plus, l’utilisation intonative de la prosodie de l’emphase-C et le lien entre les occlusives LV et l’expressivité renforcent la possibilité d’emprunter les deux (cf. Matras 2009:231 sur la position des caractéristiques prosodiques dans la hiérarchie d’emprunt en phonologie).

5.5 L’interprétation historique des faits

5.5.1 Les foyers sont des zones de rétention

Nous savons maintenant que, dans la région générale où on trouve des langues avec des occlusives LV, il y a trois foyers distincts où elles sont des phonèmes normaux, entourés de régions avec des langues dans lesquelles leur fréquence lexicale est très faible. On peut soutenir que les occlusives LV et/ou les caractéristiques phonétiques qui facilitent

leur émergence doivent avoir une plus grande profondeur temporelle dans les populations qui occupent actuellement les foyers, que dans les populations des régions environnantes. Étant donné la rareté des occlusives LV dans les langues du monde en dehors de l'Afrique subsaharienne septentrionale, il est très peu probable que les trois foyers contemporains correspondent à des zones où les occlusives LV ont émergé indépendamment. Il doit s'agir plutôt de zones actuellement occupées par des populations qui ont conservé ce trait. En revanche, les zones entourant les foyers doivent être occupées par des populations qui ont acquis des occlusives LV plus récemment, ou qui ont moins bien conservé ce trait.

La position géographique des foyers est cohérente avec leur identification comme zones de rétention. Tous les trois foyers sont des zones de refuge typiques, où le refuge doit être compris négativement dans le sens d'un lieu de dernier recours.[35] Les deux foyers occidentaux sont des zones de forêt tropicale délimitées au sud par l'océan Atlantique, à l'ouest par les chaînes de montagnes des hauts plateaux de Guinée et à l'est par les chaînes de montagnes de la ligne volcanique du Cameroun à la frontière entre le Nigeria et le Cameroun (cf. la figure 49 et la figure 50). Ils sont séparés les uns des autres par le Ghana Gap. Le Ghana Gap correspond approximativement au Dahomey Gap, une zone de savane arborée qui interrompt la forêt tropicale côtière et qui s'est établie vers 4.500 ans avant le présent, au début de l'Holocène tardif (cf. Salzmann & Hoelzmann 2005). La correspondance avec le Dahomey Gap est plus nette à la frontière orientale du foyer Upper Guinea Hotbed, où il n'y a pas d'autres frontières topographiques majeures. La frontière occidentale du foyer Lower Guinea Hotbed est délimitée par les monts Togo qui traversent le Dahomey Gap, séparant le bassin inférieur de la Volta à l'ouest d'un plateau s'inclinant progressivement vers la côte à l'est. Le Dahomey Gap a le même type de végétation et de climat que les zones à faible fréquence lexicale d'occlusives LV qui entourent les deux foyers occidentaux. Comme les deux autres foyers, la partie sud du foyer Ubangi Basin Hotbed en Afrique centrale est une zone de forêt tropicale, qui passe d'une mosaïque forêt-savane aux forêts de bas-fonds et forêts marécageuses du bassin du Congo. Sa partie nord est une sorte de cul-de-sac géographique avec de nombreuses zones marécageuses et saisonnièrement inondées. Ces deux caractéristiques écologiques contribuent à la faible densité de population que

[35] Une zone de refuge est caractérisée par des conditions environnementales qui présentent des défis de subsistance importants dont l'effet négatif est compensé par le fait que ces défis la rendent également moins attrayante pour les étrangers. Du point de vue de la dynamique des langues, les zones refuges peuvent être à la fois des zones résiduelles (ou d'accrétion) et des zones de propagation dans les termes de Nichols (1992).

j'ai déjà mise en évidence dans la figure 46 dans §5.3.3.[36] À l'ouest, le foyer Ubangi Basin hotbed est délimité par les plateaux situées à la frontière entre le Cameroun et la République centrafricaine, qui forment une extension orientale du plateau de l'Adamawa et séparent le bassin versant de l'Oubangui de celui de la Sangha, deux affluents du fleuve Congo. À l'est, ce foyer est délimité par les chaînes de montagnes de la ligne de partage Congo-Nil, aux frontières de la République centrafricaine et de la RDC avec le Soudan du Sud et l'Ouganda.

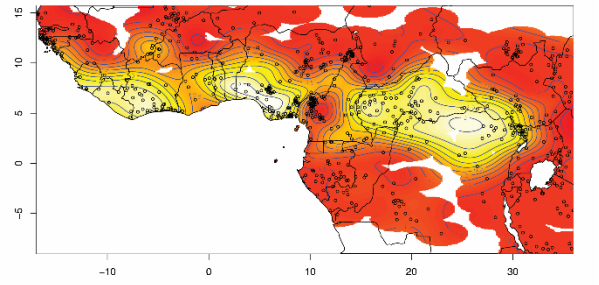


FIGURE 49. Identique à la figure 40 visualisant le GAM des valeurs de F_{LV} log-transformées.

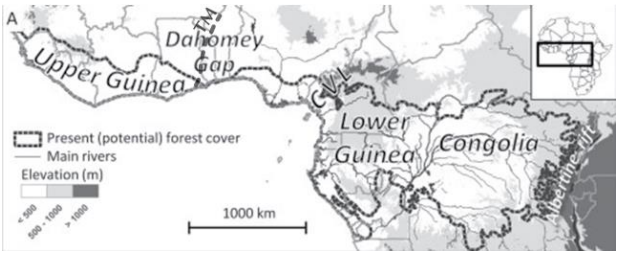


FIGURE 50. Délimitation de la forêt guinéo-congolaise, subdivision et topographie (adapté de Hardy et al. 2013). CVL : Ligne (volcanique) du Cameroun. TM : Les monts Togo.

Le scénario historique qui se dégage de ces faits est celui dans lequel des populations adaptées à la vie dans un habitat de savane et parlant des langues sans occlusives LV ont migré vers le sud, où elles ont rencontré des populations de locuteurs de langues possédant des occlusives LV et/ou les traits phonétiques qui facilitent leur apparition, que j'appellerai par souci de concision les *populations LV primaires*. La migration était la plus facile et la plus rapide dans les zones de savane. La présence marginale d'occlusives LV dans ces zones est très probablement due à un emprunt en premier lieu. Cependant, elle peut également être due à l'incorporation de plus petits sous-groupes de populations LV primaires (par exemple, par le biais de mariages mixtes)[37] et à des migrations ultérieures occasionnelles dans la direction opposée hors des zones de refuge (cf. §5.5.3). Les populations nouvellement arrivées se sont

[36] Comme discuté dans §4.5.2, en plus d'être peu peuplée, cette région d'Afrique centrale est également assez homogène sur le plan linguistique. Elle est occupée par un petit nombre de groupes linguistiques qui sont tous plutôt peu profonds. En outre, la plupart, sinon la totalité, de ces groupes sont très probablement arrivés dans cette région d'Afrique centrale relativement récemment.

[37] Cette possibilité est suggérée par exemple par certains travaux interdisciplinaires récents sur le transfert des clics des langues dites khoisan, où les clics sont des phonèmes réguliers, vers un petit groupe de langues bantu du sud-ouest de la Zambie, comme le fwe [fwee1238], où les clics sont très marginaux. Voir Pakendorf (2014:634–635) pour un résumé concis et Bostoen & Sands (2012) et Barbieri et al. (2013) pour plus de détails.

répandues beaucoup plus lentement dans les zones de refuge de la forêt tropicale, où la fréquence lexicale élevée des occlusives LV est beaucoup plus probablement le résultat d'une conversion linguistique des populations LV primaires vers les langues des nouveaux arrivants. Il s'agit d'un cas d'interférence de substrat induite par le changement de langue (« shift-induced substrate interference »), selon les termes de Thomason & Kaufman (1988).

Typiquement, les effets du substrat sur une langue cible sont directement proportionnels au ratio de la population qui a changé de langue parmi ses locuteurs (Thomason 2017). En outre, bien que les phonèmes dits empruntés puissent être transférés en même temps que les emprunts lexicaux qui les contiennent, on pense généralement que ce transfert de phonèmes est soumis à de fortes contraintes, qui ne peuvent être surmontées que par un contact intense et un haut degré de bilinguisme (cf. Winford 2003:54–56; 2005; Dimmendaal 2011:182). Sur la base d'une ligne d'argumentation similaire, Bostoen & Donzo (2013:458) soutiennent, par exemple, en ce qui concerne l'émergence des occlusives LV en ngombe, une langue bantu de la RDC, que "the integration of labial-velar stops [in Ngombe] could only happen through advanced Bantu-Ubangi bilingualism/multilingualism, probably accompanied by language shift of entire Ubangi language groups to neighbouring Bantu languages involving phonological substrate influence of their first language on the target language".

Les chances sont très élevées que les occlusives LV et/ou les caractéristiques phonétiques qui facilitent leur émergence proviennent d'une ou de plusieurs familles de langues aujourd'hui disparues. Ainsi, selon une évaluation récente de l'état de la classification des langues en Afrique par Güldemann (2018a), l'écrasante majorité des groupes linguistiques qui sont limités aux foyers sont des membres « robustes » ou « prometteurs » de la famille Niger-Congo. Le seul candidat « faible » à l'affiliation niger-congo dans les foyers est le groupe ijoïde, un petit groupe linguistique parlé dans le delta du Niger. Quelques groupes linguistiques sur les franges orientales du foyer Ubangi Basin Hotbed sont membres de la famille soudanique central. Il est important de noter que certains groupes de la même famille soudanique central et de nombreux groupes de la même famille nigéro-congo se trouvent également en dehors des zones sensibles et que ces groupes sont souvent dépourvus d'occlusives LV. Comme mentionné dans §5.3.2.3, la langue avec une fréquence lexicale très élevée d'occlusives LV qui pourrait être particulièrement intéressante dans cette région est le birri, car il s'agit d'un isolat potentiel.

Le fait que la plupart des groupes linguistiques que l'on trouve actuellement à l'intérieur et à l'extérieur des foyers en Afrique subsaharienne septentrionale sont des langues nigéro-congo indique clairement que les populations qui parlaient des langues

sans occlusives LV, qui étaient adaptées à la vie dans un habitat de savane et qui ont migré vers le sud étaient en grande partie des locuteurs de langues nigéro-congo. L'interprétation de la distribution spatiale des fréquences lexicales élevées des occlusives LV me permet de formuler des hypothèses détaillées concernant les modèles de migration préhistorique des populations parlant les langues niger-congo. A titre d'exemple, je me concentrerai dans la section suivante sur le bantoïde, un sous-groupe niger-congo de bas niveau, dont l'expansion a fait l'objet de nombreuses études.

5.5.2 Les scénarios de l'expansion bantu

Mes données sur la fréquence lexicale des occlusives LV ont des implications historiques particulièrement intéressantes pour les langues bantoïdes, un sous-groupe majeur de la branche benue-congo du niger-congo, et surtout pour son plus grand sous-groupe, les langues bantu au sens strict (Narrow Bantu), qui s'étend actuellement sur une grande partie du continent, du nord-est du Nigeria et du Kenya au nord jusqu'à l'Afrique du Sud. Avec les langues tchadiques, qui n'ont aucun rapport avec les langues bantoïdes, et un certain nombre de petits groupes des langues niger-congo apparentés, actuellement regroupés sous l'appellation des langues adamawa, les langues bantu sont les langues responsables de l'existence du Cameroon Gap entre le foyer Lower Guinea Hotbed et le foyer Ubangi Basin Hotbed. En outre, une majorité écrasante des langues bantu est parlée en dehors des foyers de fréquence lexicale élevée des occlusives LV, à l'exception notable des langues bantu du nord de la RDC et du Congo, qui font partie du foyer Ubangi Basin Hotbed. Cette distribution générale suggère que les populations bantoïdes ont dû traverser la région du Benue River Link sur leur chemin vers le sud sans beaucoup d'interaction avec les populations LV primaires. Ce passage a dû se produire quelque part dans la période comprise entre environ 4.500 avant le présent, la période présumée de la diversification initiale du bantu, et environ 6.900 avant le présent, la période présumée de la diversification initiale de l'ensemble du groupe bantoïde (cf. Grollemund et al. 2015; Bostoen et al. 2015).

On considère actuellement que la diversification initiale du bantoïde et plus tard du bantu s'est produite dans la région générale du plateau des Grassfields à l'ouest du Cameroun (cf. Greenberg 1972; Grollemund et al. 2015; Bostoen et al. 2015). Cependant, je pense qu'il est plus probable qu'elle ait commencé dans un endroit plus au nord, peut-être quelque part au nord de l'extrémité occidentale du plateau d'Adamawa, plus près des monts Alantika. La raison en est qu'au cours de l'Holocène précoce, entre environ 11.000 et 6.000 ans avant le présent, la forêt tropicale africaine s'étendait jusqu'au plateau d'Adamawa et peut-être même jusqu'à la moyenne vallée du Benue, bien au-delà des Grassfields au Cameroun, et ce n'est qu'à la fin de cet optimum

climatique que la forêt a commencé à se fragmenter sur le plateau d'Adamawa (cf. Vincens et al. 2010). Ainsi, étant donné la préférence des groupes benue-congo en général et des groupes bantoïdes en particulier pour les environnements de savane,[38] toute migration significative de ces populations plus au sud avant le début de la fragmentation de la forêt est moins probable. Cette interprétation correspond également mieux aux données paléoanthropologiques des deux seuls sites dans les Grassfields pour les périodes pertinentes de l'Holocène ancien et moyen, à savoir les abris rupestres du cratère de Mbi et de Shum Laka (Asombang 1988; Orban et al. 1996; Lavachery 2001; Lipson et al. 2019).[39]

Après leur diversification initiale, la plupart des groupes bantoïdes sont restés concentrés dans la même région générale. La principale exception est représentée par l'expansion bantu, l'un des plus grands événements d'expansion connus dans l'histoire du continent. Deux scénarios principaux pour l'expansion bantu ont été proposés dans la littérature, « l'Est issu de l'Ouest » (« East out of West ») et « l'Est parallèle à l'Ouest » (« East separate from West »), qui diffèrent selon qu'ils considèrent le bantu oriental comme une ramification ultérieure d'un nœud bantu occidental ou comme une branche primaire du nœud proto-bantu respectivement (pour un aperçu général, voir Bostoen & Grégoire 2007; Pakendorf, Bostoen & de Filippo 2011; Grollemund et al. 2015; Bostoen et al. 2015). Les routes migratoires générales associées aux deux scénarios sont illustrées dans les figures 51a et 51b, tandis que la figure 51c montre l'emplacement approximatif des principaux sous-groupes bantu mentionnés dans le texte.

[38] Ainsi, comme le concluent Grollemund et al. (2015:4) à propos de l'expansion bantu qui a suivi la diversification initiale des langues bantoïdes, « the Bantu expansion was characterized by a measureable preference for following familiar savannah habitats [and] avoided rainforest habitats » et actuellement « rainforest-dwelling Bantu cultures […] retained some cultural knowledge of how to exploit the savannah environment ».

[39] Le squelette adulte du site du cratère de Mbi daté d'environ 9.000-8.400 avant le présent a une taille comparable à celle des populations pygmées actuelles, tandis que les squelettes adultes des deux phases funéraires du site de Shum Laka datées respectivement d'environ 8.000-7.500 avant le présent et 3.900-3.000 avant le présent, varient en taille entre celle des populations pygmées actuelles et celle des populations bantu actuelles de la région. Cependant, Lipson et al. (2019) rapportent que leur analyse des données génomiques de l'ADN des individus enterrés à Shum Laka entre 8.000 et 3.000 avant le présent montre que leurs profils d'ascendance sont plus similaires à ceux des chasseurs-cueilleurs de l'ouest de l'Afrique centrale. À cet égard, la variation de la taille des individus de Shum Laka pourrait fournir des preuves supplémentaires de la possibilité, suggérée par Skoglund et al. (2017:67, e14), que la taille plus courte des populations actuelles de chasseurs-cueilleurs de la forêt tropicale soit une évolution relativement récente.

a. L'Est issu de l'Ouest

b. L'Est parallèle à l'Ouest

c. Les principaux sous-groupes bantu

FIGURE 51. (a), (b) : Les deux principaux modèles de l'expansion bantu. 1 = le nœud proto-bantu, 2 = le nœud bantu occidental, 3 = le nœud bantu oriental. (c) : La localisation approximative des principaux sous-groupes bantu : A – bantu nord-ouest, B – bantu lebonya-boan, C – bantu du bassin intérieur du Congo, D – bantu occidental côtier, E – bantu sud-ouest, F – bantu oriental (adapté de Pakendorf, Bostoen & de Filippo 2011:6, 8). La zone grise dans (a) et (b) montre le domaine bantu, tandis que dans (c) elle marque la partie de l'Afrique australe en dehors du domaine bantu. Les symboles d'arbres dans les trois figures montrent l'étendue actuelle de la forêt tropicale.

Mes données sur la fréquence lexicale des occlusives LV soutiennent clairement le modèle de l'expansion bantu « l'Est issu de l'Ouest », le point de séparation du bantu oriental se situant quelque part au sud de la forêt tropicale. Selon le modèle alternatif « l'Est parallèle à l'Ouest », les populations proto-bantu oriental auraient migré au nord de la forêt tropicale du bassin du Congo, mais cette route migratoire se trouverait en plein cœur du foyer Ubangi Basin Hotbed où la fréquence lexicale des occlusives LV est élevée, comme l'illustre la figure 52. Si tel avait été le cas, nous aurions dû trouver de nombreuses langues bantu orientales avec des occlusives LV et avec une fréquence lexicale relativement élevée de celles-ci. Cependant, les occlusives LV sont particulièrement marginales dans le bantu oriental, comme on peut facilement observer en comparant le domaine du bantu oriental dans la figure 51c (où le bantu oriental est étiqueté F) avec la visualisation GAM de la distribution spatiale des fréquences des LV dans la figure 40, reproduite ici dans la figure 52.

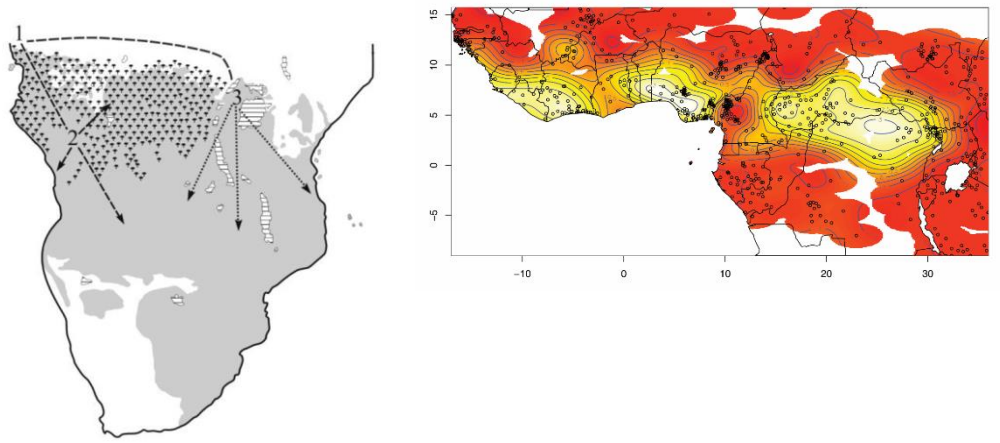
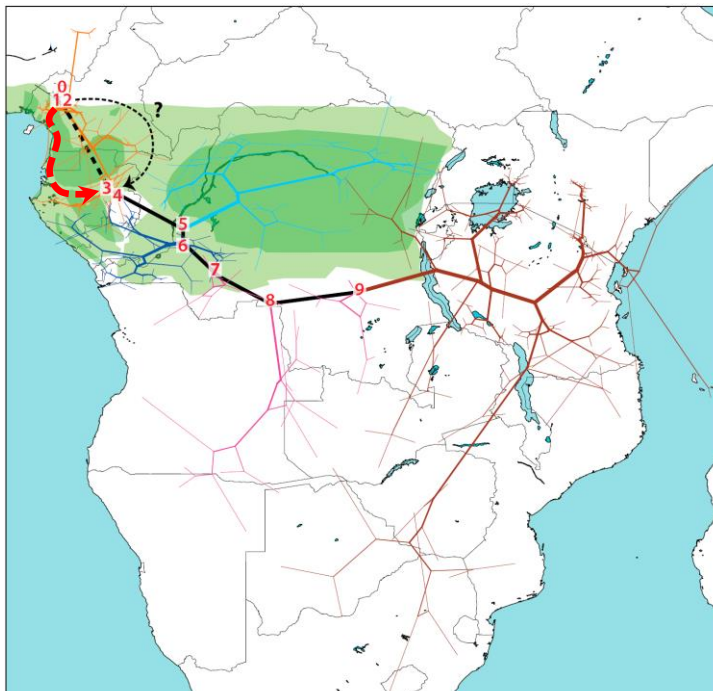


FIGURE 52. Le modèle de l'expansion bantu « l'Est issu de l'Ouest » (adapté de Pakendorf, Bostoen & de Filippo 2011:8) en comparaison avec la visualisation GAM de la distribution spatiale des fréquences des LV dans la figure 40 (le GAM avec des valeurs F_{LV} log-transformées).

Le scénario de l'expansion bantu « l'Est issu de l'Ouest » a été récemment corroboré par les résultats d'un certain nombre d'études interdisciplinaires combinant des données phylogénétiques linguistiques, de génétique des populations, archéologiques et paléoclimatiques (de Filippo et al. 2012; Bostoen et al. 2015; Grollemund et al. 2015). Grollemund et al. (2015) et Bostoen et al. (2015) proposent également une reconstruction détaillée de la route migratoire bantu, illustrée dans la figure 53a. Bostoen et al. (2015) se concentrent sur l'itinéraire que les populations bantu ont pris pour traverser pour la première fois la forêt équatoriale. Ils affirment que cette migration initiale a emprunté les corridors de savane du Sangha River Interval. Au nord de la forêt équatoriale, autour de la région de confluence Sanaga-Mbam, ces corridors de savane ont commencé à s'ouvrir vers 4.000-3.500 avant le présent, mais le passage à travers le cœur de la forêt équatoriale dans le Sangha River Interval lui-même n'a été ouvert que vers ~2.500 avant le présent. Cet itinéraire de migration supposé à travers les corridors de savane est indiqué dans la figure 53a par une flèche noire courbé en pointillés reliant les nœuds 2 et 3 et accompagnée d'un point d'interrogation. Cependant, l'itinéraire de migration à travers le Sangha River Interval ne concorde pas bien avec mes données sur la fréquence lexicale des occlusives LV (et accessoirement, avec mes données sur la prosodie de l'emphase-C), car il ferait passer les premières populations bantu par la périphérie occidentale du foyer Ubangi Basin Hotbed où la fréquence lexicale des occlusives LV est élevée (cf. la figure 53b). En conséquence, on se serait attendus à trouver un nombre significatif de langues bantu avec des occlusives LV au sud de la forêt tropicale parmi les langues bantu occidentales côtières, les langues bantu sud-ouest et les langues bantu orientales. Pourtant, cette attente ne se concrétise pas, comme on peut l'observer en comparant la figure 53a et la figure 53b. Mes propres données sur la fréquence lexicale des occlusives LV suggèrent que la migration entre

les nœuds 2 et 3 est plus susceptible de s'être effectuée à travers les savanes des plaines côtières qui se sont ouvertes déjà à partir de 4.000 avant le présent. Ce passage possible est indiqué dans la figure 53a par une flèche rouge courbée en pointillés. À cet égard, il est révélateur que les langues bantu actuellement parlées le long de la côte du sud-ouest du Cameroun et du Gabon, à savoir la plupart des langues bantu A20 et A30 et les langues bantu B10, ressemblent beaucoup plus aux langues bantu orientales dans leur phonologie et leur morphosyntaxe. Toutes sont dépourvues d'occlusives LV ou en ont une faible fréquence, et n'ont qu'une prosodie d'empase-C limitée ou n'en ont pas du tout.



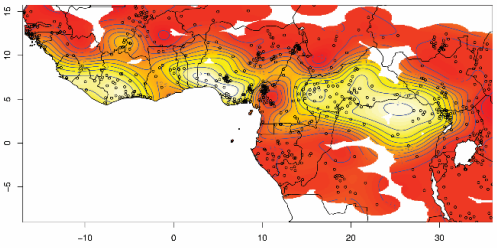


FIGURE 53b. Identique à la figure 40 visualisant le GAM des valeurs de F_{LV} log-transformées.

FIGURE 53a. La route migratoire bantu reconstruite par Grollemund et al. (2015) sur l'arbre de consensus en utilisant les emplacements géographiques des langues contemporaines et en reliant les emplacements ancestraux par des lignes droites (adapté de Grollemund et al. 2015:6). Les positions numérotées correspondent aux nœuds de diversification majeurs sur l'arbre de consensus. L'ombrage vert clair correspond à la délimitation de la forêt tropicale à 5.000 avant le présent ; le vert plus foncé correspond à la délimitation de la forêt tropicale à 2.500 avant le présent. La flèche noire courbée en pointillés indique la voie de migration à travers le Sangha River Interval proposée par Bostoen et al. (2015). La flèche rouge courbée en pointillés indique la route de migration par les savanes des plaines côtières qui correspond mieux à mes données sur la fréquence lexicale des

Du point de vue de mes données, la route passant par le Sangha River Interval était plutôt une route secondaire utilisée plus tard (à partir de ~2.500 avant le présent) par les populations ancestrales des groupes bantu A80 et A90, l'autre clade qui s'est séparé au nœud 2. Ces langues bantu ont dû être affectées par des interférences de substrat induites par des populations LV. Certains de ces groupes bantu A qui ont utilisé la route du Sangha River Interval sont susceptibles d'avoir passé par la suite aux langues bantu

occidentales côtières des groupes bantu B50-B80 parlées dans la région centrée sur les plateaux Batéké, les savanes des hautes terres à l'est du Gabon et au centre du Congo. Ceci est suggéré par la présence de la prosodie d'emphase-C (cf. §5.4) dans certaines de ces langues, comme le kukuya [teke1280], une langue B70 (Paulian 1975). Mes données sur la fréquence des occlusives LV et la prosodie d'emphase-C suggèrent en outre qu'un certain nombre d'autres groupes bantu nord-ouest sont susceptibles d'avoir initialement pris une route migratoire similaire vers l'est en direction de l'Afrique centrale, qui passait probablement un peu plus au nord le long du plateau d'Adamawa, mais qui ont ensuite soit, comme le bantu A70, fait demi-tour en direction du sud-ouest le long des franges septentrionales de la forêt tropicale, soit, comme le bantu jarawan, fait demi-tour en direction du nord-ouest le long du fleuve Benue dans le nord-est du Nigeria.[40] Bostoen et al. (2015:365) envisagent aussi brièvement la possibilité que les populations bantu aient pris une route côtière pour traverser pour la première fois la forêt équatoriale, mais la rejettent finalement en faveur du scénario de la route passant par le Sangha River Interval, faute de preuves archéologiques suffisantes sur l'existence de communautés villageoises dans les plaines côtières entre 4.000 et 3.000 avant le présent. En même temps, comme Bostoen et al. (2015:366) le reconnaissent également, aucunes données archéologiques sur l'existence de communautés villageoises n'est disponible pour le Sangha River Interval pour la période ~2.500 avant le présent non plus.

5.5.3 Les occlusives labiales-vélaires ne doivent pas être reconstruites dans le proto-niger-Congo et le proto-soudanique central

Un certain nombre de publications résumées et approuvées par Cahill (2017; 2018) soutiennent que les occlusives LV pourraient être reconstruites dans les proto-langues des principales sous-branches du niger-congo et dans le proto-niger-congo lui-même. Ces affirmations sont généralement basées sur l'observation correcte mais historiquement non pertinente que les occlusives LV peuvent être trouvées dans de nombreuses langues filles de ces familles. Cependant, la distribution géographique des fréquences lexicales élevées des occlusives LV suggère fortement qu'elles ne devraient pas être reconstruites dans la proto-langue du niger-congo ou de ses principales sous-branches, sauf dans celles qui pourraient être parlées dans l'un des foyers. Les langues

[40] En ce qui concerne la route migratoire jarawan proposée, notons que deux populations jarawan, les Nagumi [nagu1244] et les Mbonga [mbon1252], auraient été trouvées au début du 20ème siècle le long de cette route, respectivement dans le nord et l'est du Cameroun (près de la frontière avec la RCA) (cf. Maddieson & Williamson 1975). Les deux langues semblent déjà avoir disparu. Des traditions orales revendiquant une migration vers l'aval le long du Benue ont également été rapportées pour les Mbula, une autre population jarawan du nord-est du Nigeria (Meek 1931:57–68).

niger-congo se sont propagées dans les foyers à partir du nord. Plus les langues niger-congo contemporaines sont éloignées des foyers, moins elles sont susceptibles d'avoir des occlusives LV, ce qui montre qu'elles ont dû acquérir ce trait lors de leur expansion dans les foyers. Les mêmes considérations s'appliquent au soudanique central.

Des données comparatives le confirment. Dans la majorité des cas de correspondances entre une occlusive LV et une autre consonne, cette autre consonne est une occlusive vélaire qui est soit labialisée, soit suivie d'une voyelle arrondie. La reconstruction de tels ensembles de correspondances avec une occlusive LV impliquerait une lénition et la perte du relâchement labial, mais pour des raisons phonotactiques et perceptives, une telle évolution est hautement improbable. Le relâchement labial des occlusives LV est perceptiblement plus saillant (cf. Ladefoged & Maddieson 1996:336–339; Connell 1994; Cahill 2018), ce qui rend sa perte en faveur de l'articulation vélaire peu probable. Une perte généralisée du relâchement labial dans les langues en dehors des foyers est encore plus improbable parce que les occlusives LV sont typiquement limitées à la position initiale du radical, qui est souvent aussi initiale du mot. En position initiale du mot, c'est précisément le geste labial qui est le plus susceptible de masquer le geste vélaire, car dans les occlusives LV, le relâchement labial suit le relâchement vélaire. En revanche, la reconstruction de tels ensembles de correspondances avec une occlusive vélaire explique naturellement pourquoi les occlusives LV ont tendance à être restreintes à la position initiale du radical (C_1), car la proéminence prosodique de cette position facilite leur émergence (voir §5.4). Le scénario inverse nécessite la lénition des occlusives LV héritées dans l'environnement prosodique où la lénition est le moins susceptible de se produire. Pour toutes ces raisons, je suis fortement en accord avec les reconstructions qui postulent l'émergence des occlusives LV à partir des vélaires labialisées, comme Creissels (2004) pour les langues mandingues [mand1435] et Hyman (2011:13–14) pour les langues bantu, et je trouve particulièrement peu plausibles les reconstructions qui soutiennent la direction du

changement opposée, comme la reconstruction des occlusives LV pour le proto-soudanique central de Boyeldieu (2006b).[41]

Les exemples les plus convaincants de la perte des occlusives LV impliquent des ensembles de correspondances dans lesquels les LV correspondent à des consonnes labiales, en raison de la plus grande proéminence perceptive du relâchement labial dans les occlusives LV. Ils ont tendance à être plus rares, se trouvent principalement en dehors des foyers et sont généralement dus à des évolutions récentes. Dans mon scénario, la perte des occlusives LV est plus susceptible de se produire lorsque les communautés de locuteurs quittent les foyers et intègrent un nombre significatif de locuteurs de langues sans occlusives LV. L'akan [akan1250] et le supyire [supy1237] sont des exemples potentiels intéressants. Ces deux langues sont parlées en dehors des deux foyers ouest-africains et sont inhabituelles dans leurs groupes linguistiques de

[41] Boyeldieu (2006b) présente un certain nombre d'ensembles de cognats qui, selon lui, devraient être reconstruits avec les occlusives LV **k͡p*, **g͡b* et **ŋ͡m* en proto-soudanique central. Il reconstruit également ces ensembles de cognats avec des occlusives LV dans certaines branches du soudanique central, comme le sara-bongo-bagirmi. Il est intéressant de noter toutefois que tout comme en niger-congo, c'est le sous-groupe le plus septentrional des langues sara-bongo-bagirmi, à savoir le sara-bongo-bagirmi occidental, qui est généralement dépourvu d'occlusives LV (cf. figure 23 pour une carte des langues sara-bongo-bagirmi). De même, comme en niger-congo, les occlusives LV dans les langues sara-bongo-bagirmi qui en possèdent correspondent dans les langues sara-bongo-bagirmi qui n'en possèdent pas non seulement à des occlusives labiales simples mais aussi à des occlusives vélaires simples, ce qui pour les mêmes raisons que dans le cas des langues niger-congo, réduit encore la plausibilité de leur reconstruction dans le proto-sara-bongo-bagirmi en tant qu'occlusives LV. En plus de cela, le foyer d'origine des langues sara-bongo-bagirmi est également située en dehors d'un foyer actuel de haute fréquence lexicale des occlusives LV, à savoir le foyer Ubangi Basin Hotbed (cf. figure 24). Pour les langues lendu, une autre branche présumée du soudanique central, Boyeldieu (2006b) reconstruit des occlusives LV, à savoir **k͡p*, **g͡b*, **ŋ͡m*, dans certains des ensembles de cognats en question, mais des affriquées coronales dans d'autres, viz. **ʦ*, **ʣ*, **nʣ*, ce qui soulève la question de la cohérence entre les deux types d'ensembles de cognats. Pour les deux autres branches présumées du soudanique central, le moru-ma'di et le mangbetu-qsua, Boyeldieu (2006b) reconstruit les mêmes ensembles de cognats avec des vélaires labialisées, à savoir **kʷ*, **gʷ*, **ŋʷ*.

niveau inférieur en raison de l'absence d'occlusives LV (selon Cahill 2018:155, ceci est le résultat d'une fusion des occlusives LV avec les labiales dans ces langues).[42]

Un exemple encore plus spectaculaire de la corrélation entre la localisation à l'intérieur des foyers et la présence d'occlusives LV est représenté par la variété salaga du dendi [dend1243], telle que présentée par Zima (1985). Dans cette variété, les occlusives LV ont été acquises lors d'une migration dans un foyer, puis perdues lors d'une migration ultérieure en dehors du foyer au milieu du 19ème siècle. Le dendi est le membre le plus méridional de la famille songhay dont les autres membres sont parlés dans le Sahel, principalement le long du coude du Niger. Le cœur du domaine du dendi est situé au nord du foyer Lower Guinea Hotbed, à la frontière entre le Niger et le Bénin. Lorsqu'une communauté de locuteurs du dendi s'est déplacée vers le sud jusqu'à la ville de Djougou dans l'ouest du Bénin, à l'intérieur du foyer Lower Guinea Hotbed, les occlusives LV se sont développées dans cette variété de dendi par un changement phonologique régulier à partir des vélaires labialisées. Après une migration ultérieure au milieu du 19ème siècle, une partie de la communauté des Dendi de Djougou s'est installée dans la ville de Salaga au centre du Ghana, à l'intérieur du Ghana Gap de faible fréquence lexicale des occlusives LV. Dans les années 1980, les occlusives LV de la variété salaga n'étaient préservées que dans le discours des locuteurs âgés, dans un nombre limité d'expressions fixes, alors qu'en dehors de ces expressions fixes, les locuteurs de tous âges utilisent les vélaires labialisées correspondantes dans les mêmes mots. Cette inversion apparente d'un changement phonologique régulier suggère qu'au moment de la migration de Djougou à Salaga, le changement phonologique des vélaires

[42] Le supyire, une langue senufo, et l'akan, une langue potou-tano, sont tous deux situés à l'extérieur des foyers et leurs familles respectives se sont probablement propagées vers le nord à partir du foyer Upper Guinea Hotbed. Ainsi, le domaine senufo dans son ensemble est généralement orienté sud-nord, suivant en grande partie l'extension Banfora Extension du foyer Upper Guinea Hotbed (cf. §5.3.1), tandis que le supyire est également l'une des langues senufo les plus septentrionales. Le domaine potou-tano est généralement orienté sud-ouest-nord-est et s'étend du sud-est de la Côte d'Ivoire (dans le foyer Upper Guinea Hotbed) au nord du Togo et du Bénin. La zone de la plus grande diversité linguistique au sein du potou-tano (son centre de gravité) et vraisemblablement sa région d'origine se trouve dans le sud-ouest de son domaine. L'expansion vers le nord de la branche tano du potou-tano (la branche incluant l'akan) dans l'actuel Dahomey Gap est probablement liée à un retour temporaire à des conditions climatiques plus humides et à une nouvelle propagation des forêts dans la savane dans cette zone entre environ 3.300 et 1.100 avant le présent (cf. Salzmann & Hoelzmann 2005). En plus d'affecter négativement la fréquence des occlusives LV, l'expansion vers le nord du tano semble également avoir affecté négativement la fréquence des implosives dans les inventaires phonologiques de ces langues, comme le suggèrent les données de Clements & Rialland (2008:58). L'expansion du tano elle-même peut être en grande partie responsable d'une discontinuité majeure dans la distribution d'un autre trait aréal important dans l'Afrique subsaharienne septentrionale, à savoir le marquage de la négation en fin de phrase (cf. §4.5.1.2).

labialisées en occlusives LV n'était pas encore complet et qu'il y avait encore des variations dans la réalisation entre les occlusives LV et les vélaires labialisées.

5.6 Conclusions

La distribution aréale inégale des occlusives LV est connue depuis des décennies et a été incluse dans des ensembles de traits utilisés pour caractériser de grandes aires linguistiques en Afrique subsaharienne septentrionale, telles que l'aire Sudanic zone (Clements & Rialland 2008) ou l'aire Macro-Sudan Belt (Güldemann 2008). Le développement récent de la grande base de données lexicales RefLex m'a permis d'aller au-delà de la simple énumération des langues qui ont des occlusives LV dans leur inventaire de phonèmes et d'estimer le degré d'ancrage lexical de ces phonèmes dans un très large échantillon de 315 langues qui en possèdent. Comme je l'avais prévu, sur la base de mes connaissance des langues individuelles, les occlusives LV se sont avérées être des phonèmes marginaux dans une proportion non négligeable de mon échantillon. Lorsque j'ai ensuite étudié la distribution géographique des fréquences lexicales des occlusives LV, une distribution inattendue et très intéressante est apparue : trois foyers de fréquence lexicale élevée de LV entourés de zones de fréquence lexicale faible de LV. Ces foyers recoupent les groupes généalogiques et sont plus facilement caractérisés en termes géographiques : ils ont une faible altitude et des habitats forestiers ou marécageux et ils sont séparés les uns des autres par des zones d'altitude plus élevée et/ou un habitat de savane. Ce profil détaillé indique clairement que l'origine et la distribution actuelle des occlusives LV sont un phénomène de substrat : un trait des communautés linguistiques qui ont trouvé refuge dans les foyers au moment de l'expansion des langues niger-congo et des langues soudaniques centrales vers le sud et qui se sont finalement converties aux langues des groupes entrants. Ce scénario implique que les occlusives LV ne faisaient pas partie des inventaires de phonèmes des proto-langues des familles actuellement attestées en Afrique subsaharienne septentrionale, ce qui est en accord avec les arguments purement linguistiques en termes de naturalité perceptive et phonotactique de l'évolution des vélaires labialisées vers des labiales-vélaires en position initiale de mot ou de racine, plutôt que l'évolution inverse.

Mes données sur la distribution spatiale des fréquences lexicales élevées des occlusives LV m'ont également permis de formuler des hypothèses détaillées concernant les routes migratoires préhistoriques des populations niger-congo. En particulier, j'ai pu ajuster et affiner les scénarios proposés dans la littérature pour l'expansion bantu, l'un des plus grands événements d'expansion linguistique de l'histoire humaine récente.

Finalement, mes données quantitatives m'ont permis de renforcer la base empirique de l'affirmation selon laquelle les occlusives LV sont plus fréquentes dans les parties expressives du vocabulaire que dans le vocabulaire général. Je l'ai fait indirectement en montrant qu'elles sont moins fréquentes dans le vocabulaire de base, non expressif, des listes Swadesh 200 que dans le vocabulaire général. Ce fait distributif a une explication commune avec le fait qui pourrait sembler sans rapport que les occlusives LV sont prépondérantes en position initiale du radical. Je soutiens que ces deux phénomènes sont dus à la prosodie d'emphase-C, la proéminence prosodique des consonnes initiales du radical dont le corrélat phonétique typique est la longueur de la consonne. Mes recherches en cours montrent que la prosodie d'emphase-C est utilisée au niveau de l'énoncé dans de nombreuses langues de l'Afrique subsaharienne septentrionale pour marquer une emphase particulière sur un élément donné.

Références

Andersen, Torben. 1981. *A grammar of Modo: A preliminary sketch*. Aalborg: University Centre of Aalborg.

Asombang, Raymond Neb'ane. 1988. *Bamenda in prehistory: The evidence from Fiye Nkwi, Mbi Crater and Shum Laka rockshelters*. London: University of London PhD thesis.

Baayen, R. Harald. 2013. Multivariate statistics. In Robert J. Podesva & Devyani Sharma (eds.), *Research methods in linguistics*, 337–372. Cambridge: Cambridge University Press.

Baddeley, Adrian & Rolf Turner. 2005. spatstat: An R package for analyzing spatial point patterns. *Journal of Statistical Software* 12(6). 1–42.

Barbieri, Chiara, Anne Butthof, Koen Bostoen & Brigitte Pakendorf. 2013. Genetic perspectives on the origin of clicks in Bantu languages from southwestern Zambia. *European Journal of Human Genetics* 21. 430–436.

Beyer, Klaus. 2009. Double negation marking: A case of contact-induced grammaticalization in West-Africa? In Norbert Cyffer, Erwin Ebermann & Georg Ziegelmeyer (eds.), *Negation patterns in West African languages and beyond*, 205–222. Amsterdam: John Benjamins.

Bisang, Walter & Remi Sonaiya. 2000. Information structuring in Yoruba. *Linguistics* 38(1). doi: 10.1515/ling.38.1.169.

Blench, Roger & Kay Williamson. 2016. A reconstruction of the phonology of Proto-Igboid. URL: http://llacan.cnrs.fr/nigercongo2/discussions/Proto-Igboid_phonology.pdf.

Bostoen, Koen, Bernard Clist, Charles Doumenge, Rebecca Grollemund, Jean-Marie Hombert, Joseph Koni Muluwa & Jean Maley. 2015. Middle to Late Holocene paleoclimatic change and the early Bantu expansion in the rain forests of western Central Africa. *Current Anthropology* 56(3). 354–384. doi: 10.1086/681436.

Bostoen, Koen & Jean-Pierre Donzo. 2013. Bantu-Ubangi language contact and the origin of labial-velar stops in Lingombe (Bantu, C41, DRC). *Diachronica* 30(4). 435–468.

Bostoen, Koen & Claire Grégoire. 2007. La question bantoue: bilan et perspectives. *Mémoires de la Société de Linguistique de Paris* 15. 73–91.

Bostoen, Koen & Bonny Sands. 2012. Clicks in south-western Bantu languages: contact-induced vs. language-internal lexical change. In Matthias Brenzinger (ed.), *Proceedings of the 6th World Congress of African Linguistics Cologne 2009*, 129–140. Köln: Rüdiger Köppe.

Boyeldieu, Pascal. 2006a. Présentation des langues Sara-Bongo-Baguirmiennes. CNRS-LLACAN. URL: https://llacan.cnrs.fr/SBB/Boyeldieu_SBB.pdf.

Boyeldieu, Pascal. 2006b. Reflexes of a labiovelar series in Central Sudanic. In Al-Amin Abu-Manga, Leoma Gilley & Anne Storch (eds.), *Insights into Nilo-Saharan Language, History and Culture: Proceedings of the 9th Nilo-Saharan Linguistic Colloquium, Institute of African and Asian Studies, University of Khartoum, 16-19 February 2004*, vol. 23, 129–151. (Nilo-Saharan). Köln: Rüdiger Köppe. URL: https://halshs.archives-ouvertes.fr/halshs-00331321.

Boyeldieu, Pascal & France Cloarec-Heiss. 1986. Dialectometrie lexicale dans le domaine oubanguien. In Gladys Guarisma & Wilhelm J. G. Möhlig (eds.), *La méthode dialectometrique appliquée aux languages africaines*, 331–393. Berlin: Dietrich Reimer Verlag.

Brown, Dunstan, Marina Chumakina & Greville G. Corbett (eds.). 2013. *Canonical morphology and syntax*. Oxford: Oxford University Press.

Bybee, Joan & Shelece Easterday. 2019. Consonant strengthening: A crosslinguistic survey and articulatory proposal. *Linguistic Typology* 23(2). 263–302. doi: 10.1515/lingty-2019-0015.

Cahill, Michael. 2008. Why labial-velar stops merge to /gb/. *Phonology* 25(3). 379–398. doi: 10.1017/S0952675708001541.

Cahill, Michael. 2017. Labial-velars: A questionable diagnostic for a linguistic area. In Shigeki Kaji (ed.), *Proceedings of the 8th World Congress of African Linguistics*, 13–24. Tokyo: Research Institute for Languages and Cultures of Asia and Africa, Tokyo University of Foreign Studies.

Cahill, Michael. 2018. Labial-velars of Africa: Phonetics, phonology, and historical development. In Augustine Agwuele & Adams Bodomo (eds.), *The Routledge handbook of African linguistics*, 150–167. Abingdon, Oxon: Routledge. doi: 10.4324/9781315392981.

Campbell, Lyle. 2006. Areal linguistics: A closer scrutiny. In Yaron Matras, April McMahon & Nigel Vincent (eds.), *Linguistic areas: Convergence in historical and typological perspective*, 1–31. Hampshire: Palgrave Macmillan. doi: 10.1057/9780230287617_1.

Clements, G. N. & Annie Rialland. 2008. Africa as a phonological area. In Bernd Heine & Derek Nurse (eds.), *A linguistic geography of Africa*, 36–85. Cambridge: Cambridge University Press.

Connell, Bruce. 1991. *Phonetic aspects of the Lower Cross languages and their implications for sound change*. Edinburgh: University of Edinburgh PhD thesis.

Connell, Bruce. 1994. The structure of labial-velar stops. *Journal of Phonetics* 22(4). 441–476.

Crane, Thera M., Larry M. Hyman & Simon Nsielanga Tukumu. 2011. *A grammar of Nzadi [B865]: A Bantu language of Democratic Republic of Congo*. Berkeley: University of California Press.

Creissels, Denis. 2004. L'occlusive vélaire sonore *g* et les labio-vélaires (*w, gw, kw, gb, kp*) en mandingue. *Mandenkan* 39. 1–22.

Creissels, Denis. 2005. S-O-V-X constituent order and constituent order alternations in West African languages. In Rebecca Cover & Yuni Kim (eds.), *Proceedings of the Berkeley Linguistics Society 31st annual meeting*, 37–51. Berkeley: University of California at Berkeley.

Cyffer, Norbert, Erwin Ebermann & Georg Ziegelmeyer. 2009. *Negation patterns in West African languages and beyond*. Amsterdam: John Benjamins.

de Filippo, Cesare, Koen Bostoen, Mark Stoneking & Brigitte Pakendorf. 2012. Bringing together linguistic and genetic evidence to test the Bantu expansion. *Proceedings of the Royal Society B: Biological Sciences* 279(1741). 3256–3263. doi: 10.1098/rspb.2012.0318.

Devos, Maud & Johan van der Auwera. 2013. Jespersen cycles in Bantu: double and triple negation. *Journal of African Languages and Linguistics* 34(2). 205–274. doi: https://doi.org/10.1515/jall-2013-0008.

Dimmendaal, Gerrit J. 2011. *Historical linguistics and the comparative study of African languages*. Amsterdam: John Benjamins.

Dockum, Rikker & Claire Bowern. 2019. Swadesh lists are not long enough: Drawing phonological generalizations from limited data. *Language Documentation and Description* 16. 35–54.

Dryer, Matthew S. 2006. Functionalism and the metalanguage - Theory confusion. In Grace Wiebe, Gary Libben, Tom Priestly, Ron Smyth & Sam Wang (eds.), *Phonology, morphology, and the empirical imperative: Papers in honour of Bruce Derwing*, 27–59. Taipei: The Crane Publishing Company.

Dryer, Matthew S. 2009. Negation patterns in West African languages and beyond. In Norbert Cyffer, Erwin Ebermann & Georg Ziegelmeyer (eds.), *Verb-object-negative order in Central Africa*, 307–362. Amsterdam: John Benjamins.

Fasiolo, Matteo, Raphael Nedellec, Yannig Goude & Simon N. Wood. 2018. Scalable visualisation methods for modern Generalized Additive Models. *Arxiv preprint*. URL: https://arxiv.org/abs/1707.03307.

Fisch, Maria. 1984. Die Kavangofischer. *Namibiana* 5(1). 105–169.

GeoNames.org. GeoNames database dump. URL: http://download.geonames.org/export/dump/ (Accessed on 14 March, 2016).

Greenberg, Joseph H. 1959. Africa as a linguistic area. In William R. Bascom & Melville J. Herskovits (eds.), *Continuity and change in African cultures*, 15–27. Chicago: University of Chicago Press.

Greenberg, Joseph H. 1963. *The languages of Africa*. Bloomington: Indiana University Press.

Greenberg, Joseph H. 1972. Linguistic evidence regarding Bantu origins. *Journal of African Languages and Linguistics* 13. 189–216.

Greenberg, Joseph H. 1983. Some areal characteristics of African languages. In Ivan R. Dihoff (ed.), *Current approaches to African linguistics*, vol. 1, 3–21. Dordrecht: Foris Publications.

Grollemund, Rebecca, Simon Branford, Koen Bostoen, Andrew Meade, Chris Venditti & Mark Pagel. 2015. Bantu expansion shows that habitat alters the route and pace of human dispersals. *Proceedings of the National Academy of Sciences* 112(43). 13296–13301. doi: 10.1073/pnas.1503793112.

Güldemann, Tom. 1999. The genesis of verbal negation in Bantu and its dependency on functional features and clause types. In Jean-Marie Hombert & Larry M. Hyman (eds.), *Bantu historical linguistics: theoretical and empirical linguistics*, 545–587. Stanford: CSLI.

Güldemann, Tom. 2003. Logophoricity in Africa: An attempt to explain and evaluate the significance of its modern distribution. *Sprachtypologie und Universalienforschung* 56. 366–387.

Güldemann, Tom. 2008. The Macro-Sudan belt: towards identifying a linguistic area in northern sub-Saharan Africa. In Bernd Heine & Derek Nurse (eds.), *A linguistic geography of Africa*, 151–185. Cambridge: Cambridge University Press.

Güldemann, Tom. 2010. Sprachraum and geography: Linguistic macro-areas in Africa. In Alfred Lameli, Ronald Kehrein & Stefan Rabanus (eds.), *Language and space. An international handbook of linguistic variation*, vol. 2: Language mapping, 561–585; maps 2901–2914. Berlin: Mouton de Gruyter.

Güldemann, Tom. 2018a. Historical linguistics and genealogical language classification in Africa. In Tom Güldemann (ed.), *The languages and linguistics of Africa*, 58–444. Berlin: De Gruyter Mouton. doi: 10.1515/9783110421668-002.

Güldemann, Tom. 2018b. Areal linguistics beyond contact, and linguistic areas in Afrabia. In Tom Güldemann (ed.), *The languages and linguistics of Africa*, 448–545. Berlin: De Gruyter Mouton. doi: 10.1515/9783110421668-002.

Güldemann, Tom & Tjerk Hagemeijer. 2015. How to become a Macro-Sudan belt language: The Gulf-of-Guinea creole (GGC) case. Paper presented at the Workshop “Areal phenomena in northern sub-Saharan Africa” at WOCAL 8, Kyoto, Japan. URL: http://idiatov.mardi.myds.me/Areal_Phenomena_in_NSSA.html.

Hammarström, Harald. 2018. A survey of African languages. In Tom Güldemann (ed.), *The languages and linguistics of Africa*, 1–57. Berlin: De Gruyter Mouton. doi: 10.1515/9783110421668-001.

Hammarström, Harald, Robert Forkel, Martin Haspelmath & Sebastian Bank. 2022. *Glottolog 4.6*. Leipzig: Max Planck Institute for the Science of Human History. URL: http://glottolog.org. doi: 10.5281/zenodo.6578297.

Heath, Jeffrey. 2008. *A grammar of Jamsay*. Berlin, New York: Mouton de Gruyter. doi: 10.1515/9783110207224.

Heine, Bernd. 1975. Language typology and convergence areas in Africa. *Linguistics* 144. 27–47.

Heine, Bernd & Derek Nurse (eds.). 2008. *A linguistic geography of Africa*. Cambridge: Cambridge University Press.

Hombert, Jean-Marie & Gérard Philippson. 2009. The linguistic importance of language isolates: the African case. In Peter Austin, Oliver Bond, Monik Charette, David Nathan & Peter Sells (eds.), *Proceedings of Conference on Language Documentation and Linguistic Theory 2*. London: SOAS. URL: www.hrelp.org/eprints/ldlt2_15.pdf.

Hulstaert, Gustaaf. 1957. *Dictionnaire lɔmɔ́ngɔ-français*. 2 vols. Tervuren: Musée royal du Congo belge.

Hulstaert, Gustaaf. 1961. *Grammaire du lɔmɔ́ngɔ*. Vol. 1: La phonologie. Tervuren: Musée royal de l’Afrique centrale.

Hulstaert, Gustaaf. 1965. *Grammaire du lɔmɔ́ngɔ*. Vol. 2: La morphologie. Tervuren: Musée royal de l’Afrique centrale.

Hulstaert, Gustaaf. 1966. *Grammaire du lɔmɔ́ngɔ*. Vol. 3: La syntaxe. Tervuren: Musée royal de l’Afrique centrale.

Hyman, Larry M. 2004. How to become a “Kwa” verb. *Journal of West African Languages* 30(2). 69–88.

Hyman, Larry M. 2011. The Macro-Sudan belt and Niger-Congo reconstruction. *Language Dynamics and Change* 1(1). 3–49.

Idiatov, Dmitry. 2008. Antigrammaticalization, antimorphologization and the case of Tura. In Elena Seoane, María José López-Couso & (in collaboration with) Teresa Fanego (eds.), *Theoretical and empirical issues in grammaticalization*, 151–169. Amsterdam: John Benjamins. doi: 10.1075/tsl.77.09idi.

Idiatov, Dmitry. 2012a. On the history of clause-final negation in the Mande languages of the Bani – upper Mouhoun rivers area. Paper presented at the Workshop "The history of post-verbal negation in African languages" (7th World Congress of African Linguistics). URL: http://idiatov.mardi.myds.me/WOCAL7_Negation/IDIATOV_2012_Presentation.pdf.

Idiatov, Dmitry. 2012b. Clause-final negative markers in southeastern Bamana dialects: a contact-induced evolution. *Africana Linguistica* 18. 169–192.

Idiatov, Dmitry. 2015. Clause-final negative markers in Bobo and Samogo: parallel evolution and contact. *Journal of Historical Linguistics* 5(2). 235–266. doi: 10.1075/jhl.5.2.02idi.

Idiatov, Dmitry. 2018. An areal typology of clause-final negation in Africa: language dynamics in space and time. In Daniël Van Olmen, Tanja Mortelmans & Frank Brisard (eds.), *Aspects of linguistic variation*, 115–163. Berlin: De Gruyter Mouton. URL: https://doi.org/10.1515/9783110607963-005.

Idiatov, Dmitry. in prep. Clause-final negation in northern sub-Saharan Africa: right periphery, intersubjectivity and areality.

Idiatov, Dmitry & Mark L.O. Van de Velde. 2016. Stem-initial accent and C-emphasis prosody in north-western Bantu. Paper presented at the 6th International Conference on Bantu Languages, Helsinki, Finland. URL: http://idiatov.mardi.myds.me/talks/2016_BANTU6_C-emphasis_Idiatov_Van_de_Velde_SLIDES.pdf.

Idiatov, Dmitry & Mark L.O. Van de Velde. 2021. The lexical distribution of labial-velar stops is a window into the linguistic prehistory of Northern Sub-Saharan Africa. *Language* 97(1). 72–107. doi: 10.1353/lan.2021.0013.

Jones, Ross McCallum. 1998. *The Boko/Busa language cluster*. München: LINCOM Europa.

Kamba Muzenga, J. G. 1981. *Les formes verbales négatives dans les langues bantoues*. Tervuren: Musée royal de l'Afrique centrale.

Kleinewillinghöfer, Ulrich. 2001. Jalaa, an almost forgotten language of northeastern Nigeria: A language isolate? *Sprache und Geschichte in Afrika* 16–17. 239–271.

Ladefoged, Peter & Ian Maddieson. 1996. *The sounds of the world's languages*. Oxford: Blackwell.

Larochette, Joseph. 1959. Overeenkomst tussen Mangbetu, Zande en Bantu-talen. *Handelingen van het XXIIIe Vlaams Filologencongres*, 247–248. Brussel.

Lavachery, Philippe. 2001. The Holocene archaeological sequence of Shum Laka rock shelter (Grassfields, Western Cameroon). *African Archaeological Review* 18(4). 213–247. doi: 10.1023/A:1013114008855.

Lionnet, Florian & Larry M. Hyman. 2018. Current issues in African phonology. In Tom Güldemann (ed.), *The languages and linguistics of Africa*, 602–708. Berlin: De Gruyter Mouton.

Lipson, Mark, Mary Prendergast, Isabelle Ribot, Carles Lalueza-Fox & David Reich. 2019. Ancient Human DNA from Shum Laka (Cameroon) in the Context of African Population History. Paper presented at the Paper presented at The 84th Annual Meeting of the Society for American Archaeology, Albuquerque, NM. URL: https://core.tdar.org/document/452234/ancient-human-dna-from-shum-laka-cameroon-in-the-context-of-african-population-history.

Littig, Sabine & Ulrich Kleinewillinghöfer. 2012. Negation patterns in Sama-Duru languages (Central Adamawa). Paper presented at the Workshop "The history of post-verbal negation in African languages" (7th World Congress of African Linguistics), Buea. URL: http://idiatov.mardi.myds.me/WOCAL7_Negation/LITTIG_KLEINEWILLINGHOEFER_2012_Presentation.pdf.

Lucas, Christopher. 2009. *The development of negation in Arabic and Afro-Asiatic*. Cambridge: University of Cambridge PhD thesis.

Maddieson, Ian. 2011. Presence of uncommon consonants. In Matthew Dryer & Martin Haspelmath (eds.), *The world atlas of language structures online*. Munich: Max Planck Digital Library. URL: http://wals.info/feature/19A.

Maddieson, Ian. 2018. Phonetics and African languages. In Tom Güldemann (ed.), *The languages and linguistics of Africa*, 546–601. Berlin: De Gruyter Mouton.

Maddieson, Ian & Kay Williamson. 1975. Jarawan Bantu. *African Languages / Langues Africaines* 1. 124–163.

Maniacky, Jacky. 2003. *Tonologie du ngangela: variété de Menongue (Angola)*. München: LINCOM.

Martin, Marieke. 2015. Wawa ideophone phonetics vs. Wawa phonology. Paper presented at the World Conference of African Linguistics 8, Kyoto.

Matras, Yaron. 2009. *Language contact*. Cambridge: Cambridge University Press.

Meek, Charles K. 1931. *Tribal studies in Northern Nigeria*. . 2 vols. London: Kegan Paul, Trench, Trubner & Co.

Meeussen, Achille E. 1975. Possible linguistic Africanisms. *Language Sciences* 35. 1–5.

Miestamo, Matti. 2008. *Standard negation: The negation of declarative verbal main clauses in a typological perspective*. Berlin: Walter De Gruyter.

Moñino, Yves. 1988. Introduction: Cousines ou voisines? In Yves Moñino (ed.), *Lexique comparatif des langues oubanguiennes*, 11–22. Paris: Geuthner.

Moñino, Yves. 2004. Prête-moi ta langue, que je dise un mot: emprunts banda au gbaya. In Pierre Nougayrol & Pascal Boyeldieu (eds.), *Langues et cultures: terrains d'Afrique (hommage à France Cloarec-Heiss)*, 25–31. Louvain: Peeters.

Moran, Steven & Daniel McCloy (eds.). 2019. *PHOIBLE 2.0*. Jena: Max Planck Institute for the Science of Human History. URL: http://phoible.org.

Moran, Steven, Daniel McCloy & Richard Wright (eds.). 2014. *PHOIBLE Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. URL: http://phoible.org.

Nichols, Johanna. 1992. *Linguistic diversity in space and time*. Chicago: University of Chicago Press.

Nigeria Federal Surveys. 1958–1973. Nigeria 1:100.000. Lagos: Nigeria Federal Surveys.

Olson, Kenneth S. & John Hajek. 2003. Crosslinguistic insights on the labial flap. *Linguistic Typology* 7(2). 157–186. doi: 10.1515/lity.2003.014.

Orban, Rosine, Isabelle Ribot, Sylvie Fenaux & Pierre de Maret. 1996. Les restes humains de Shum Laka (Cameroun, LSA-Age du fer). *Anthropologie et Préhistoire* 107. 213–225.

Pakendorf, Brigitte. 2014. Historical linguistics and molecular anthropology. In Claire Bowern & Bethwyn Evans (eds.), *The Routledge handbook of historical linguistics*, 627–641. London: Routledge. URL: https://halshs.archives-ouvertes.fr/halshs-01179242.

Pakendorf, Brigitte, Koen Bostoen & Cesare de Filippo. 2011. Molecular perspectives on the Bantu expansion: A synthesis. *Language Dynamics and Change* 1(1). 50–88.

Paulian, Christiane. 1975. *Le kukuya: langue teke du Congo*. Paris: SELAF.

Persson, Andrew M. & Janet R. Persson. 1991. *Mödö-English dictionary with grammar*. Nairobi: Summer Institute of Linguistics.

R Core Team. 2015. R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing. URL: http://www.R-project.org.

Rolle, Nicholas, Florian Lionnet & Matthew Faytak. 2020. Areal patterns in the vowel systems of the Macro-Sudan Belt. *Linguistic Typology* 24(1). 113–179. doi: 10.1515/lingty-2019-0028.

Rombi, Marie-Françoise & Jacqueline M. C. Thomas. 2006. *Un continuum prédicatif: Le cas du Gbanzili (République Centrafricaine)*. Louvain, Paris: Peeters.

RStudio Team. 2016. RStudio: Integrated development for R. Boston, MA: RStudio, Inc. URL: http://www.rstudio.com.

Salzmann, Ulrich & Philipp Hoelzmann. 2005. The Dahomey Gap: An abrupt climatically induced rain forest fragmentation in West Africa during the late Holocene. *Holocene* 15(2). 190–199. doi: 10.1191/0959683605hl799rp.

Scholz, Hans-Jürgen. 1976. *Igbira phonology*. Dallas: SIL.

Segerer, Guillaume & Sébastien Flavier. 2011–2022. *RefLex: Reference Lexicon of Africa*. Version 2. Paris, Lyon: CNRS. URL: http://reflex.cnrs.fr.

Shimizu, Kiyoshi. 1979. *A comparative study of the Mumuye dialects (Nigeria)*. Berlin: Reimer.

Skoglund, Pontus, Jessica C. Thompson, Mary E. Prendergast, Alissa Mittnik, Kendra Sirak, Mateja Hajdinjak, Tasneem Salie, et al. 2017. Reconstructing prehistoric African population structure. *Cell* 171(1). 59-71.e21. doi: 10.1016/j.cell.2017.08.049.

Solomiac, Paul. 2007. *Phonologie et morphosyntaxe du dzùngoo de Samogohiri*. Lyon: Université Lumière Lyon 2 PhD thesis.

Swadesh, Morris. 1952. Lexicostatistic dating of prehistoric ethnic contacts. *Proceedings American Philosophical Society* 96. 452–463.

Tamminga, Meredith, Christopher Ahern & Aaron Ecay. 2016. Generalized Additive Mixed Models for intraspeaker variation. *Linguistics Vanguard* 2(s1). 1–9. doi: 10.1515/lingvan-2016-0030.

Thomason, Sarah G. 2017. On establishing ancient shift-induced interference: Problems and prospects. Paper presented at the Workshop "Language shift and substratum interference in (pre)history," Max Planck Institute for the Science of Human History, Jena.

Thomason, Sarah G. & Terrence Kaufman. 1988. *Language contact, creolization, and genetic linguistics*. Berkeley: University of California Press.

Tisserant, Charles. 1930. *Essai sur la grammaire banda*. Paris: Institut d'Ethnologie.

Tosco, Mauro. 2001. *The Dhaasanac language: Grammar, texts, vocabulary of a Cushitic language of Ethiopia*. Cologne: Rüdiger Köppe.

Van de Velde, Mark L.O. 2008. *A grammar of Eton*. Berlin: Mouton de Gruyter.

Van de Velde, Mark L.O. 2012. The origin and spread of possessee-like qualifiers in Central Africa. Paper presented at the 7th World Congress of African Languages, Buea. URL: http://llacan.cnrs.fr/pers/vandevelde/files/Van_de_Velde_WOCAL_BUEA.pptx .

Van de Velde, Mark L.O. 2013. The Bantu connective construction. In Anne Carlier & Jean-Christophe Verstraete (eds.), *The genitive*, 217–252. Amsterdam: John Benjamins.

Van de Velde, Mark L.O. to appear. Approaches to comparative Bantu morphosyntax. *Linguistique et Langues Africaines*.

van der Auwera, Johan. 2009. The Jespersen cycles. In Elly van Gelderen (ed.), *Cyclical change*, 35–71. Amsterdam: John Benjamins.

van der Auwera, Johan. 2010. On the diachrony of negation. In Laurence R. Horn (ed.), *The expression of negation*, 73–101. Berlin: Mouton de Gruyter.

van der Auwera, Johan & Lauren Van Alsenoy. 2016. On the typology of negative concord. *Studies in Language* 40(3). 473–512. doi: 10.1075/sl.40.3.01van.

van Gelderen, Elly. 2008. Negative cycles. *Linguistic Typology* 12(2). 195–243. doi: 10.1515/LITY.2008.037.

Veselinova, Ljuba. 2013. Negative existentials: a cross-linguistic study. *Rivista di Linguistica* 25(1). 107–145.

Vincens, A., G. Buchet, M. Servant & ECOFIT Mbalang collaborators. 2010. Vegetation response to the "African Humid Period" termination in Central Cameroon (7° N) – new pollen insight from Lake Mbalang. *Climate of the Past* 6(3). 281–294. doi: 10.5194/cp-6-281-2010.

Vogler, Pierre. 2014. La formation des labiales-vélaires à double occlusion en Niger-Congo. URL: https://hal.archives-ouvertes.fr/hal-01183115/document.

Westermann, Diedrich. 1911. *Die Sudansprachen: eine sprachvergleichende studie*. Hamburg: L. Friederichsen.

Westermann, Diedrich. 1927. *Die westlichen Sudansprachen und ihre Beziehungen zum Bantu*. Berlin: De Gruyter.

Wieling, Martijn. 2018. Analyzing dynamic phonetic data using generalized additive mixed modeling: A tutorial focusing on articulatory differences between L1 and L2 speakers of English. *Journal of Phonetics* 70. 86–116. doi: 10.1016/j.wocn.2018.03.002.

Wieling, Martijn, Simonetta Montemagni, John Nerbonne & R. Harald Baayen. 2014. Lexical differences between Tuscan dialects and Standard Italian: Accounting for geographic and sociodemographic variation using Generalized Additive Mixed Modeling. *Language* 90(3). 669–692.

Wieling, Martijn, John Nerbonne & R. Harald Baayen. 2011. Quantitative social dialectology: Explaining linguistic variation geographically and socially. *PLoS ONE* 6(9). e23613. doi: 10.1371/journal.pone.0023613.

Winford, Donald. 2003. *An introduction to contact linguistics*. Malden MA: Blackwell.

Winford, Donald. 2005. Contact-induced changes: Classification and processes. *Diachronica* 22(2). 373–427.

Winkelmann, Kerstin & Gudrun Miehe. 2009. Negation in Gur: Genetic, areal and unique features. In Norbert Cyffer, Erwin Ebermann & Georg Ziegelmeyer (eds.), *Negation patterns in West African languages and beyond*, 167–204. Amsterdam: John Benjamins.

Winter, Bodo & Martijn Wieling. 2016. How to analyze linguistic change using mixed models, Growth Curve Analysis and Generalized Additive Modeling. *Journal of Language Evolution* 1(1). 7–18. doi: 10.1093/jole/lzv003.

Wood, Simon N. 2006. *Generalized Additive Models: An introduction with R*. Boca Raton: Chapman and Hall–CRC.

Wood, Simon N. 2019. mgcv: Mixed GAM Computation Vehicle with Automatic Smoothness Estimation. URL: http://CRAN.R-project.org/package=mgcv.

Zima, Petr. 1985. Labiovelar stops in the Djougou Dendi dialect of Songhay. *Acta Universitatis Carolinae - Philologica 3: Phonetica pragensia* VII. 97–104.