La fréquence lexicale des occlusives labiales-vélaires dans le nord de l'Afrique sub-saharienne

Dmitry Idiatov, Mark van de Velde

▶ **To cite this version:**

Dmitry Idiatov, Mark van de Velde. La fréquence lexicale des occlusives labiales-vélaires dans le nord de l'Afrique sub-saharienne. Jean-Léo Léonard & Annie Rialland. Linguistique africaine: perspectives croisées, Peeters, pp.189-204, 2018. halshs-01956334

HAL Id: halshs-01956334
https://shs.hal.science/halshs-01956334

Submitted on 15 Dec 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

La fréquence lexicale des occlusives labiales-vélaires dans le nord de l'Afrique sub-saharienne

Dmitry IDIATOV & Mark VAN DE VELDE

Abstract

Cross-linguistically, labial-velar stops are rather rare, but they are known to be common in the phonological inventories of many genealogically diverse languages spoken in northern sub-Saharan Africa. Using quantitative data, this paper shows that the distribution of the lexical frequency of labial-velar stops is very uneven in the languages of this area. First, we can distinguish between languages with low frequency of labial-velar stops in the lexicon and languages with high lexical frequency of labial-velar stops. Importantly, the spatial distribution of the two groups is not random. There are three hotbeds of high lexical frequency of labial-velar stops and a peripheral area of low lexical frequency. Second, we show that within individual languages significant differences in the frequencies of labial-velar stops exist between the so-called basic vocabulary and the lexicon as the whole, which we take as indirect evidence for the hypothesis that labial-velar stops are more frequent in the "expressive" parts of the lexicon, such as ideophones.

1. Introduction[1]

Il a été remarqué depuis longtemps que le nord de l'Afrique sub-saharienne (NASS) constitue une zone de diffusion de traits linguistiques. Parmi les travaux récents qui essayent de prendre une perspective englobante sur les phénomènes aréaux dans le NASS, il faut mentionner les hypothèses d'aréalité de Güldemann (2008), qui parle de l'aire macro-Soudan (« Macro-Sudan belt ») qu'il détermine sur la base de six critères morphosyntaxiques et phonologiques, et de Clements & Rialland (2008), qui parlent de la zone soudanique (« Sudanic zone ») et qui ne considèrent {p. 190} que des critères phonologiques. Dans le cadre de nos propres recherches sur les phénomènes aréaux dans le NASS, on s'intéresse aux traits qui ont une distribution aréale marquée et aux manières d'expliquer leur émergence et diffusion, ainsi qu'aux manières d'évaluer leur ancienneté et leur stabilité dans la perspective de leur éventuelle pertinence pour la reconstruction des proto-langues.

[1] Cette recherche s'insère dans le projet LC2 "Areal phenomena in Northern sub-Saharan Africa" of the Labex EFL (financé par une aide de l'Etat gérée par l'Agence Nationale de la Recherche au titre du programme « Investissements d'Avenir » portant la référence ANR-10-LABX-0083).

L'un des traits linguistiques dont nous savons depuis longtemps qu'il a une fréquence importante dans les langues du NASS tout en étant très rare ailleurs est la présence des occlusives labiales-vélaires, notamment /k͡p/, /g͡b/ et /ŋ͡m/ (cf. Cahill 2008, Maddieson 2011). Pour cette raison, la présence des occlusives labiales-vélaires (LV) dans les inventaires phonologiques est toujours mentionnée parmi les traits aréaux distinctifs du NASS, quelle que soit la délimitation exacte de l'aire linguistique en question. En même temps, un examen rapide des descriptions des langues qui ont des LV révèle que ces langues peuvent varier considérablement en ce qui concerne le statut de ces consonnes dans leurs phonologies et lexiques (voir, par exemple, Bostoen & Donzo 2013, qui examinent de près le statut des LV dans quelques langues bantu et oubangiennes du nord de la RDC). Afin de mieux évaluer le statut des consonnes LV dans les langues du NASS, notamment en vue d'une étude ultérieure sur leur origine et diffusion, nous avons quantifié cette variation dans un échantillon de 336 langues à consonnes labiales-vélaires. Pour ce faire, nous avons comparé pour chaque langue la fréquence lexicale attestée des LV à la fréquence attendue dans la situation canonique où chaque consonne de la langue aurait la même fréquence lexicale. Les résultats de cette étude quantitative confirment les observations plus partielles ou impressionnistes sur le statut souvent marginal des occlusives LV dans les langues en question.

Si en moyenne les consonnes LV ont une faible fréquence lexicale dans les langues du NASS, il n'en est pas de même pour toutes les langues de la région et nous avons trouvé que la fréquence lexicale relative des LV varie en fonction de la position géographique des langues. Nous avons effectué une analyse statistique de la répartition spatiale des fréquences lexicales des LV dans le NASS, qui a démontré l'existence de deux foyers de haute fréquence lexicale. Un premier foyer se situe le long de la côte Atlantique entre le Libéria et le Nigeria. Le deuxième se situe plus à l'est en Centrafrique et dans le nord de la RDC. Ces deux régions sont séparées par une discontinuité majeure au niveau du Cameroun et du nord-est du Nigeria. La première région, la zone côtière de l'Afrique de l'ouest, est scindée en deux sous-régions par une discontinuité moins fortement prononcée au niveau du Ghana.

Finalement, nous avons voulu vérifier l'hypothèse selon laquelle les occlusives labiales-vélaires seraient plus fréquentes dans le lexique dit « expressif », ce qui correspond plus ou moins aux idéophones et aux qualifiants évaluatifs. Malheureusement, {p. 191} les informations nécessaires pour pouvoir extraire les fréquences des LV dans les parties expressives des lexiques des langues de notre échantillon sont absentes de nos sources de données. Faute de manière directe pour vérifier cette hypothèse, nous l'avons fait de façon indirecte en comparant la fréquence relative des LV dans une partie du lexique de base non-expressive à la fréquence relative des LV dans le lexique complet. Nous avons trouvé que la fréquence des LV dans le

lexique de base est moins élevée et que la différence est significative. Si cette approche indirecte est valide, l'hypothèse de départ se voit donc confirmée.

Dans la Section 2, nous présentons notre échantillon des langues et nos sources de données. La Section 3 procède à une évaluation quantitative du statut des LV dans les phonologies des langues du NASS. Nous y comparons la fréquence lexicale observée des LV à leur fréquence attendue dans la situation canonique où toutes les consonnes auraient la même fréquence. Dans la Section 4, nous effectuons des analyses supplémentaires pour évaluer la répartition des LV au sein des lexiques des langues du NASS, en essayant de voir si la fréquence des LV est significativement différente dans le domaine du lexique dit « de base » en comparaison avec le lexique général. La Section 5, finalement, propose une analyse statistique de la répartition spatiale des fréquences lexicales des LV dans le NASS. Elle commence avec une visualisation des résultats sous forme d'un simple graphique d'interpolation spatiale en 5.1 suivi d'une modélisation et visualisation à l'aide de l'outil de modèles additifs généralisés (GAM) dans la Section 5.2.

2. Les données

Notre principale source de données est la base de données RefLex (Segerer & Flavier 2011-2016), qui recueille plus d'un millier de sources lexicales sur les langues africaines accompagnées d'un bon nombre d'outils d'exploitation. Nous avons écarté les sources parues avant 1900, qui posent souvent d'importants problèmes de fiabilité par rapport à la transcription des LV. De plus, nous avons écarté toutes les sources qui ont moins que 100 entrées. Ce seuil a été choisi de façon arbitraire afin de pouvoir obtenir des données plus ou moins représentatives pour une langue sans exclure les listes de mots du genre lexique de base, qui constituent les seules sources de données lexicales pour beaucoup de langues africaines. Finalement, si RefLex contient deux sources pour la même langue, nous avons utilisé celle qui est plus grande et/ou de meilleure qualité. Nous avons complété ces données avec les lexiques de quelques langues mandé et bantu qui ne sont pas (encore) représentées dans RefLex. Finalement, nous avons pris en compte les informations sur la présence {p. 192} ou l'absence des LV dans les langues africaines disponibles dans la base de données Phoible (www.phoible.org).

Notre échantillon compte 1304 langues au total, dont 566 langues avec LV dans leurs inventaires phonologiques et 738 langues sans LV. Des 566 langues avec LV, il y en a 336 dont nous disposons de données sur la fréquence lexicale des LV. La Fig. 1 montre la répartition

géographique des 1304 langues de notre échantillon.[2] De plus, elle représente cette répartition sous forme d'intensité spatiale, c.-à-d. le degré de concentration des langues prises comme des points dans l'espace. On peut facilement discerner trois régions de forte concentration des langues. La plus importante se situe au niveau de la frontière entre le Cameroun et le Nigeria. La deuxième occupe une grande partie de l'Afrique de l'est et est centrée autour du lac Victoria. La troisième se trouve à l'extrême ouest du NASS.

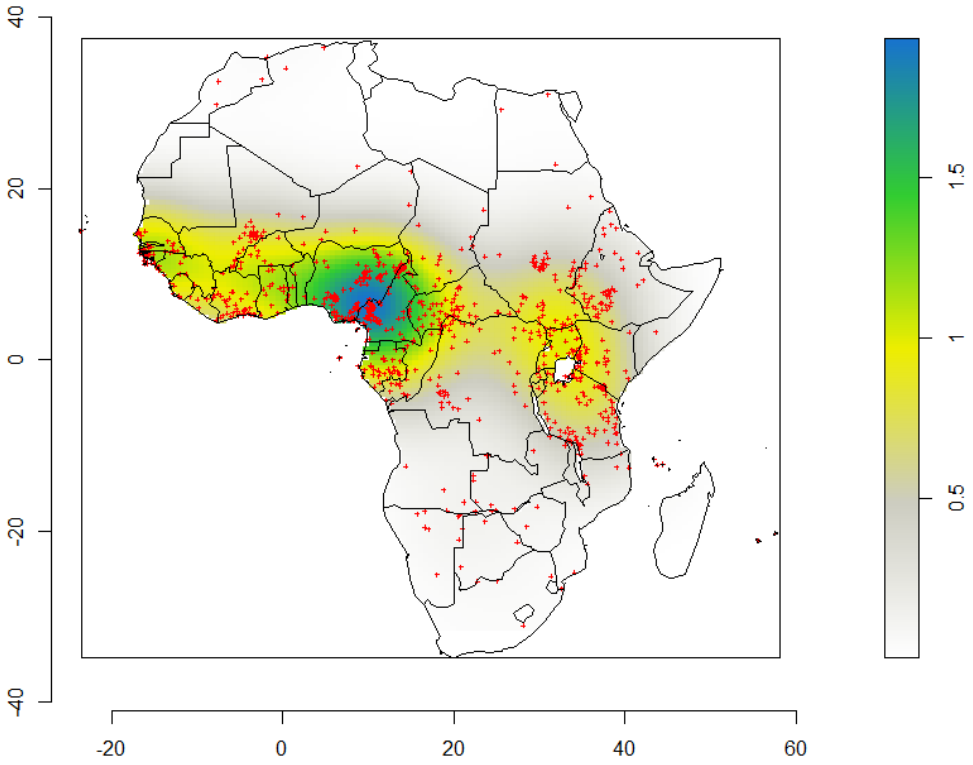


Figure 1. La répartition géographique des 1304 langues de notre échantillon et son intensité spatiale

{p. 193} La Fig. 2 montre, sous forme d'intensité spatiale, la répartition géographique des langues qui ont des LV (Fig. 2a) et des langues qui n'en ont pas (Fig. 2b). La répartition des langues avec des LV dans la Fig. 2a correspond évidemment assez bien à la zone soudanique de Clements & Rialland (2008) et à l'aire Macro-Soudan de Güldemann (2008). De plus, on peut observer que les langues avec LV sont particulièrement concentrées dans le sud du Nigeria sur un axe est-ouest. En même temps, les langues sans LV sont particulièrement concentrées aux extrêmes ouest et est du NASS. Il y a aussi une concentration assez importante des langues sans LV dans le nord-est du Nigeria et le sud-ouest du Cameroun qui chevauche partiellement la zone

[2] Tous les graphiques et les calculs dans cet article sont produits avec le logiciel *R* (R Core Team 2015). Les graphiques de l'intensité spatiale et de l'interpolation spatiale sont produits avec le paquet *spatstat* (Baddeley & Turner 2005).

avec la plus forte concentration des langues avec des LV, ce qui s'explique par la fragmentation linguistique extrême dans cette sous-région. Toutefois, il est important de noter que la zone de concentration des langues sans LV a une orientation spatiale différente, nommément sur un axe nord-sud, et qu'elle est située un peu plus vers l'est. Nous serons confrontés à une configuration similaire quand nous considèrerons de près la répartition spatiale de la fréquence lexicale des LV dans la Section 5.

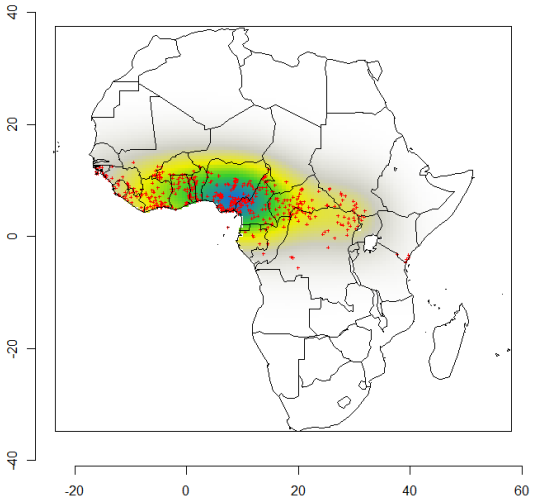


Figure 2a. La répartition géographique des 566 langues avec LV et son intensité spatiale

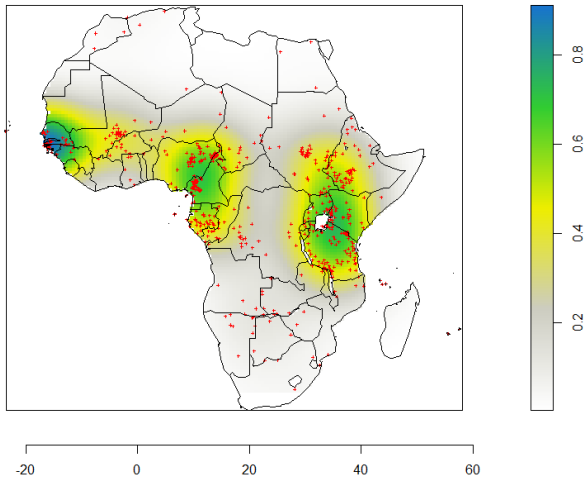


Figure 2b. La répartition géographique des 738 langues sans LV et son intensité spatiale

Avant de procéder aux calculs, nous avons nettoyé les données moissonnées dans RefLex pour les normaliser et pour enlever les erreurs occasionnelles. A part les erreurs ponctuelles dans l'encodage des consonnes dans RefLex, nous avons dû également écarter un nombre de langues où les séquences graphiques de vélaire et labiale, tels que *kp* ou *gb*, ne représentent pas des LV mais des séquences d'une consonne vélaire et une consonne labiale. Par la suite, nous avons recodé les digraphes non reconnus en tant que tels par RefLex. Finalement, nous avons scindé {p. 194} les clusters de consonnes que RefLex compte par défaut comme des unités, à savoir les soi-disant « prénasalisées » (des combinaisons d'une nasale homorganique et une consonne orale), tels que *nd* ou *mb*, les suites des consonnes avec des marques de labialisation, comme *bw*, de palatalisation, comme *by*, et les suites formées d'une consonne labiale et une consonne labiodentale, comme *bv*.

3. L'estimation de la fréquence lexicale des LV dans les langues du NASS

En tant qu'hypothèse de départ H_0 pour l'estimation de la fréquence lexicale des LV nous partons de la situation canonique dans laquelle toutes les consonnes ont la même probabilité d'occurrence

dans le lexique d'une langue donnée de notre échantillon (1). Une situation canonique est un point de départ théorique à partir duquel on peut mesurer la variation, qui n'est pas forcément attesté dans le monde réel.

(1) **H_0** : Dans le lexique d'une langue de notre échantillon, toutes les consonnes ont la même probabilité d'occurrence (c.-à-d. la même fréquence)

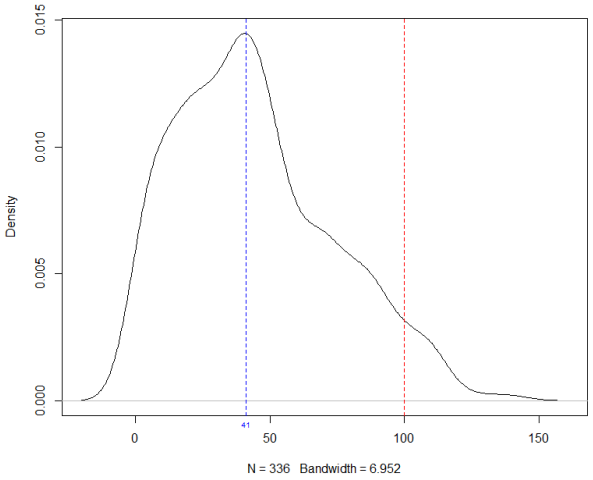
Nous avons estimé la fréquence des LV dans une langue L (F_{LV}) selon la formule présentée en (2).

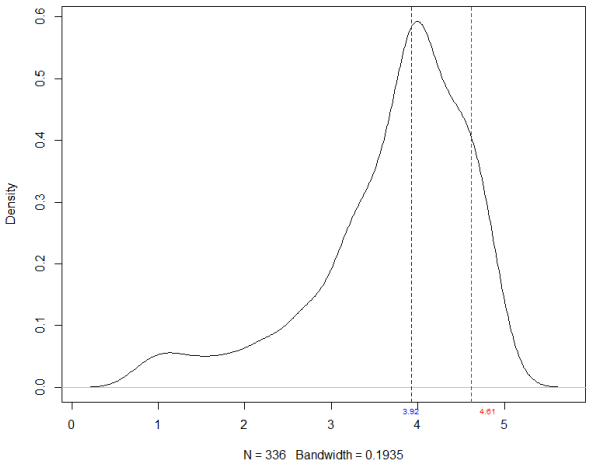
(2) $$F_{LV} = \frac{LV_O}{LV_E} * 100\% = \frac{\sum T_{LV}}{\frac{\sum T_C}{\sum P_C} * \sum P_{LV}} * 100\%$$

La fréquence des labiales-vélaires (F_{LV}) dans le lexique d'une langue donnée, exprimée en pourcentage, est la proportion du nombre de LV observées dans le lexique (LV_O) par rapport au nombre attendu (LV_E) dans la situation canonique H_0. Ce dernier est calculé en multipliant le quotient du nombre de consonnes observées dans la source lexicale ($\sum T_C$) par le nombre de phonèmes consonantiques de la langue $\sum P_C$ avec le nombre de phonèmes consonantiques labiales-vélaires $\sum P_{LV}$.

Calculée de cette manière, une F_{LV} de 0% correspond à l'absence des LV dans la langue en question, tandis qu'une F_{LV} égale à 100% implique que le nombre observé des occurrences des LV dans la langue en question est celle qui serait attendu étant donnée la H_0, ou en d'autres termes, que F_{LV} est « normale » étant donné la H_0. Nous appelons cette dernière fréquence (F_{LV} égale à 100%) la F_{LV} de référence.

Les résultats du calcul des F_{LV} dans les 336 langues du NASS de notre échantillon qui possèdent des LV sont synthétisés dans la Fig. 3 sous forme d'un graphe {p. 195} représentant la densité de probabilité de F_{LV}. Dans la Fig. 3a, les F_{LV} (en abscisse) sont des pourcentages, tandis que dans la Fig. 3b les F_{LV} sont log-transformées. Les valeurs log-transformées ont été mises à l'échelle en ajoutant la valeur F_{LV} minimale différente de zéro. La transformation log a été utilisée pour essayer de réduire l'influence d'éventuelles valeurs aberrantes et de rendre les données plus normales, puisque beaucoup de tests statistiques requièrent que les données suivent une distribution normale.





la médiane

F_{LV} de référence

Figure 3a. La densité de probabilité de F_{LV} (en pourcentage)

la médiane

F_{LV} de référence

Figure 3b. La densité de probabilité de F_{LV} (log-transformée et mise à l'échelle)

Dans la Fig. 3, la médiane de la distribution des F_{LV} est largement inférieure à la F_{LV} de référence. En d'autres termes, la Fig. 3 montre que les LV sont des phonèmes relativement rares dans la majorité des langues où on les trouve.

4. La répartition des LV au sein du lexique

Pour évaluer la répartition des LV au sein du lexique, on prend comme l'hypothèse de départ H_0 que la distribution des LV au sein du lexique dans les langues de notre échantillon est aléatoire (2). Cette hypothèse de départ est parfaitement parallèle à notre hypothèse de départ pour l'estimation de la fréquence lexicale des LV (1) qui présume que dans le lexique d'une langue de notre échantillon toutes les consonnes ont la même probabilité d'occurrence. {p. 196}

(2) **H_0** : La distribution des LV au sein du lexique dans les langues de notre échantillon est aléatoire.

De diverses considérations indépendantes suggèrent que cette H_0 est probablement fausse. Tout d'abord, ceci est suggéré par le fait que les LV sont typologiquement rares et qu'elles ont également tendance à être relativement rares dans les langues qui les possèdent (cf. Section 3). De plus, un examen rapide des descriptions des langues avec des LV révèle que les LV ont tendance à être plus fréquentes dans certaines parties du lexique plutôt que dans d'autres. Par exemple, Bostoen & Donzo (2013) qui considèrent de près le statut des LV dans plusieurs langues bantu et oubangiennes du nord de la RDC constatent que dans les langues bantu en

question, où les LV sont pourtant assez fréquentes, la fréquence des LV est significativement plus haute dans la partie expressive du lexique, surtout dans les idéophones et les mots dérivés à partir des idéophones, que dans le lexique général. Martin (2015) signale pour le wawa, une langue mambiloïde parlée au Cameroun à la frontière avec le Nigeria, que bien que les LV soient rares dans le lexique général, elles ont une fréquence élevée dans les idéophones.

Dans cette perspective, notre hypothèse alternative H_A à l'hypothèse de départ H_0 est que la distribution des LV au sein du lexique dans les langues de notre échantillon n'est pas aléatoire (3). En particulier, nous supposons que les LV sont plus fréquentes en dehors du domaine du lexique dit « de base » (H_{A1}). En outre, nous supposons que les LV sont plus fréquentes dans les parties expressives du lexique, telles que les idéophones et qualifiants évaluatifs (H_{A2}).

(3) **H_A** : La distribution des LV au sein du lexique dans les langues de notre échantillon n'est pas aléatoire.

H_{A1} : Les LV sont plus fréquentes en dehors du domaine du lexique dit « de base ».

H_{A2} : Les LV sont plus fréquentes dans les parties expressives du lexique.

Pour tester notre hypothèse alternative H_{A1} que les LV sont plus fréquentes en dehors du domaine du lexique dit « de base », nous avons extrait de chaque source de taille suffisamment grande un sous-ensemble d'entrées relevant du lexique de base afin de comparer ensuite la répartition des fréquences des LV dans l'échantillon original, F_{LV}^{O}, avec celle dans l'échantillon des lexiques de base, F_{LV}^{B}. Plus exactement, pour toutes les sources à partir de 400 entrées, nous avons créé de manière automatisée des listes de 200 entrées visant à approximer la liste de Swadesh de {p. 197} 200 entrées en tant que modèle de lexique de base.[3] Les éventuelles lacunes suite à l'absence de certains sens de la liste de « Swadesh 200 » dans la source originale ont été remplies avec d'autres entrées de la source d'une manière aléatoire. Ce procédé automatisé produit donc des sous-ensembles d'entrées quasi-« Swadesh 200 ».

[3] Nous sommes reconnaissants à Benoît Legouy pour son aide avec l'automatisation de ce procédé.

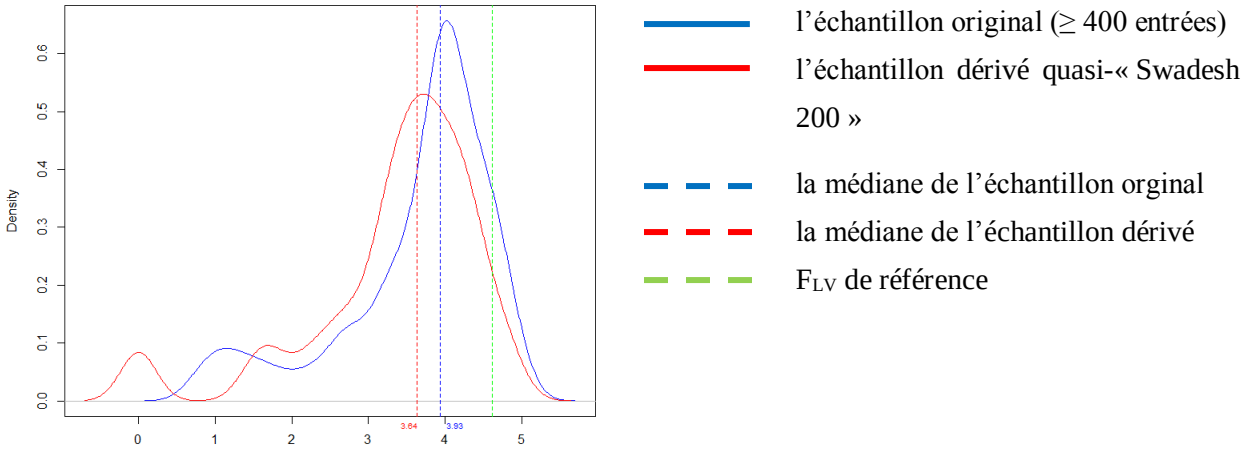


Figure 4. Les densités de probabilité de F_{LV} (log-transformées et mises à l'échelle) dans l'échantillon des sources à partir de 400 entrées et dans l'échantillon des sous-ensembles d'entrées quasi-« Swadesh 200 » dérivé à partir de ce premier

La Fig. 4 synthétise les résultats du test de notre hypothèse alternative H_{A1}. Elle présente le graphe de la densité de probabilité des F_{LV} log-transformées dans l'échantillon des sources à partir de 400 entrées (comparable à l'échantillon original dans la Fig. 3b) et y superpose le graph de la densité de probabilité des F_{LV} log-transformées dans l'échantillon dérivé des sous-ensembles d'entrées quasi-« Swadesh 200 ». Comme on peut bien l'observer la valeur médiane de F_{LV} dans l'échantillon dérivé des sous-ensembles d'entrées quasi-« Swadesh 200 » est inférieure à la valeur médiane de F_{LV} dans l'échantillon des sources complètes, qui elle-même est déjà largement inférieure à la F_{LV} de référence, ce qui suggère que les LV sont encore plus rares dans le lexique de base que dans le lexique général. Les deux distributions ne sont pas normales mais leurs variances sont comparables ce qui permet d'utiliser le test de comparaison des données de Wilcoxon (signed-rank). Ce test {p. 198} confirme qu'il est très peu probable que les deux distributions représentent la même population ($p = 5.061e\text{-}13$) et que la différence entre les moyennes de F_{LV} de ces deux jeux de données est significative. Nous avons aussi fait une validation par bootstrap (répétitions = 999) qui confirme également ce résultat (100% des valeurs $p < 0.5$, 50% des valeurs $p \leq p_0 = 5.061e\text{-}13$).

Notre test a donc confirmé notre hypothèse alternative H_{A1} que les LV sont plus fréquentes en dehors du domaine du lexique dit « de base ». Quant à notre hypothèse alternative H_{A2} selon laquelle les LV seraient plus fréquentes dans les parties expressives du lexique, elle est beaucoup plus compliquée à tester pour toutes les langues de notre échantillon, parce que pour le faire, on devrait ajouter manuellement pour chaque source les informations sur l'appartenance de chaque lexème au lexique expressif. Toutefois, même sans qu'on puisse la quantifier, la tendance générale pour les LV à être plus fréquentes dans les idéophones que dans le lexique général est

suffisamment claire, comme nous l'avons illustré plus haut à propos du wawa et des langues bantu du nord de la RDC et comme un examen rapide de beaucoup de descriptions des langues avec des LV le confirme également.

5. La répartition spatiale des fréquences lexicales des LV

Les répartitions géographiques des langues de notre échantillon présentées dans les Fig. 1 et 2 dans la Section 2 sont essentiellement représentées par des motifs de points. En tant que tels, ils nous montrent l'étendue générale des langues avec et sans LV et ils mettent en évidence les régions avec des fortes vs. faibles concentrations de ces deux types de langues. Toutefois, la question qui nous intéresse le plus, n'est pas simplement la géographie de la présence vs. l'absence des LV, relativement bien connue aujourd'hui, mais plutôt les éventuelles tendances géographiques dans la distribution des fréquences des LV dans les langues du NASS. Pour révéler de telles tendances, nous avons couplé les données de la répartition géographique des langues avec et sans LV avec les résultats de notre calcul des fréquences des LV. En tant que première approche à l'analyse de ces résultats, on propose une visualisation des résultats sous forme d'un graphique d'interpolation spatiale dans la Section 5.1. Ensuite, afin de quantifier les résultats d'une manière plus stricte, on les modélise et visualise à l'aide de l'outil de modèles additifs généralisés (GAM) dans la Section 5.2. {p. 199}

5.1. L'interpolation spatiale

Le résultat du couplage des données de la répartition géographique des langues avec et sans LV avec les résultats de notre calcul des fréquences des LV peut être inspecté visuellement dans la Fig. 5 à l'aide d'un graphique d'interpolation spatiale des fréquences log-transformées des LV dans les langues de notre échantillon.

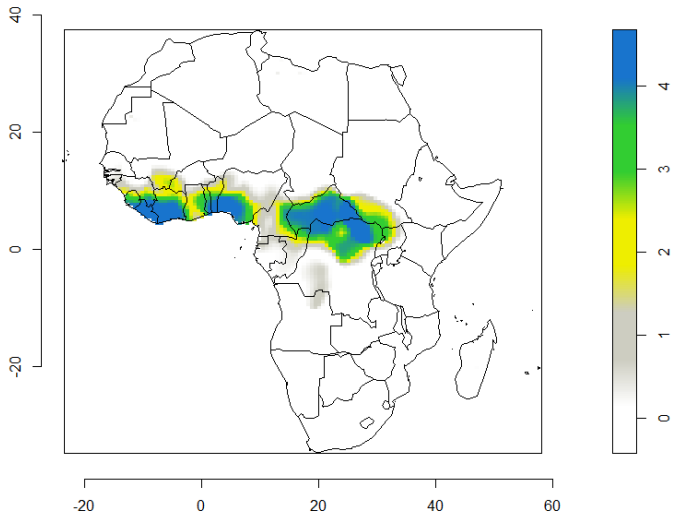


Figure 5. Le graphique d'interpolation spatiale des fréquences F_{LV} (log-transformées et mises à l'échelle) dans l'échantillon des 1304 langues africaines (l'interpolation par noyau gaussien, la valeur par défaut de la fenêtre d'estimation de la densité de probabilité ajustée de 1.5)

Sur le graphique dans la Fig. 5, on peut bien distinguer deux grandes régions avec une haute fréquence lexicale des LV, qui correspondent à peu près à la côte de l'Afrique de l'ouest, d'un côté, et la Centrafrique et le nord de la RDC, de l'autre. Ces deux régions sont séparées par une discontinuité majeure au niveau du Cameroun et du nord-est du Nigeria. En outre, la première région, la zone côtière de l'Afrique de l'ouest, se voit pratiquement scindée en deux sous-régions par une discontinuité moins fortement prononcée au niveau du Ghana. Il est également possible de distinguer sur le graphique un cluster moins saillant dans le sud-est du Mali et le sud-ouest du Burkina-Faso avec des langues à fréquence lexicale des LV relativement haute, mais qui n'atteigne jamais le niveau de la F_{LV} de référence. Ce dernier cluster pourrait également être considéré comme une sorte d'appendice de la région côtière. Finalement, on voit une sorte de pont de langues à fréquence {p. 200} lexicale des LV basse qui relie la région centrafricaine et la région côtière par la vallée de la Bénoué et le plateau d'Adamawa.

Le graphique dans la Fig. 5 est très intéressant parce qu'il montre bien que la répartition des fréquences lexicales des LV dans les langues du NASS n'est pas du tout homogène, mais se caractérise par une structure interne beaucoup plus complexe qu'on aurait pu soupçonner sur la base de la répartition géographique des langues avec des LV et son intensité spatiale présentées dans la Fig. 2a (Section 2). Bien que cette structure se prête à un nombre d'interprétations intéressantes, essayons d'abord de quantifier les résultats de ce couplage des données de la répartition géographiques des langues avec et sans LV avec les résultats de notre calcul des fréquences des LV d'une manière plus stricte.

5.2. Des modèles additifs généralisés (GAM)

Un outil statistique qui est particulièrement bien adapté à nos données est fourni par les modèles additifs généralisés (GAM). A l'origine, GAM est une extension de la régression multiple qui permet de modéliser d'une manière flexible des interactions complexes décrivant des surfaces ondulées. Pour une bonne introduction aux GAMs dans le contexte de la linguistique, on peut consulter Baayen (2013) et Winter & Wieling (2016). De bons exemples d'utilisation des GAMs dans la linguistique en relation avec l'analyse spatiale sont fournis par Wieling et al. (2011, 2014). En plus de leur capacité de traiter les données fortement non-linéaires, un grand avantage des GAMs est que c'est un outil qui laisse les données complexes parler pour elles-mêmes sans qu'on doive les recoder ou grouper en classes d'abord. Toutefois, cette liberté que les GAMs offrent et leur capacité de traitement des données très complexes ont également un certain effet secondaire. Ainsi, les GAMs ne fournissent pas des coefficients qu'on pourrait facilement interpréter d'une façon directe et leur visualisation est très importante pour leur évaluation. En outre, bien que par défaut la modélisation GAM utilise la fonction gaussienne, dans notre cas particulier, la distribution non-normale (à savoir, à queue lourde) des fréquences des LV justifie

également la modélisation GAM avec la fonction *scaled-T*. De ce fait, on va considérer deux modèles GAM, l'un produit avec la fonction gaussienne et l'autre avec la fonction *scaled-T*. Les deux GAM estiment les fréquences lexicales des LV log-transformées en fonction de la combinaison de longitude et latitude en utilisant la méthode de *thin-plate splines*, ou grilles de déformations, comme implémentée dans le paquet *mgcv* pour *R* (Wood 2006, 2015). {p. 201}

Le graphique dans la Fig. 6 représente la surface de régression du GAM produit en utilisant la fonction gaussienne sous forme d'un tracé de contours avec le schème des couleurs d'une carte thermique. Sur une carte thermique, des teints plus légers correspondent aux températures plus élevées, ce qui dans notre cas représente des fréquences lexicales des LV log-transformées plus hautes. Les lignes de contours sont des isoplèthes qui marquent les écarts de la moyenne en termes de l'écart type.

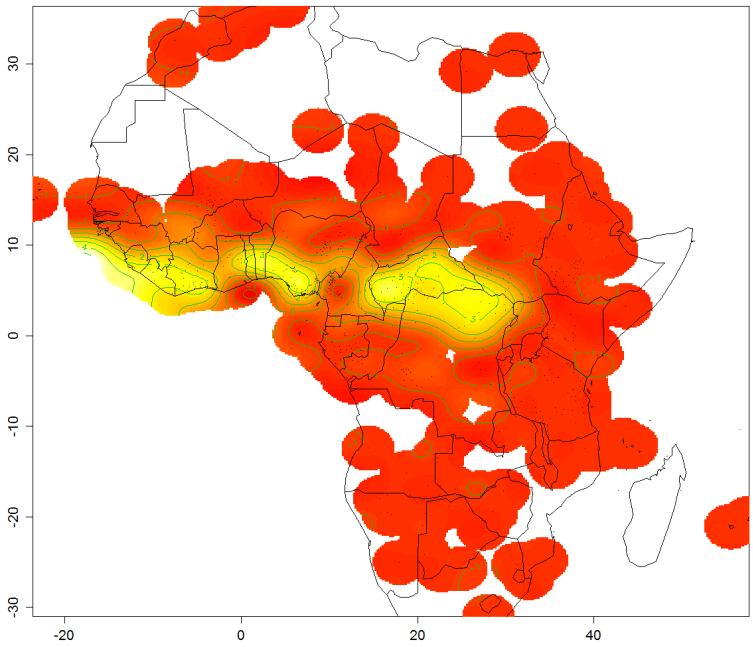


Figure 6. Le tracé de contours produit avec un modèle additif généralisé (k = 17, fonction = gaussienne, edf = 110, p < 2e-16, la déviance expliquée = 83.3%, AIC = 2602). Le tracé de contours montre la surface de régression des fréquences lexicales des LV log-transformées en fonction de la combinaison de longitude et latitude en utilisant la méthode de *thin-plate splines* avec le schème des couleurs d'une carte thermique. Les teintes les plus claires correspondent aux fréquences lexicales des LV log-transformées les plus hautes. Les lignes de contours sont des isoglosses qui marquent les écarts de la moyenne en termes d'écart type.

La visualisation du GAM dans la Fig. 6 est largement comparable au simple graphique d'interpolation spatiale dans la Fig. 5 et elle se prête aux observations similaires sur la répartition spatiale des fréquences log-transformées des LV dans les langues de notre échantillon qui ont

déjà été formulées ci-dessus. La visualisation du GAM indique toutefois une structure spatiale plus nuancée. Son apport le plus important est une meilleure mise en évidence d'une décomposition de la zone {p. 202} côtière de l'Afrique de l'Ouest en deux sous-régions du fait d'une discontinuité au niveau du Ghana. Etant donné que dans sa partie sud cette discontinuité correspond principalement à l'aire de diffusion de l'akan, une grande langue qui n'a pas de LV, on aurait pu supposer que cette discontinuité n'est qu'apparente, puisque due au hasard de la présence d'une grande langue sans LV. Toutefois, la Fig. 6 montre bien l'existence d'un bon nombre de langues à faible fréquence lexicale des LV parlées au nord de l'akan, qui contribuent à l'émergence de cette discontinuité.

Le graphique dans la Fig. 7 représente la surface de régression du GAM produit en utilisant la fonction *scaled-T* sous forme d'un tracé de contours avec le schème des couleurs d'une carte thermique.

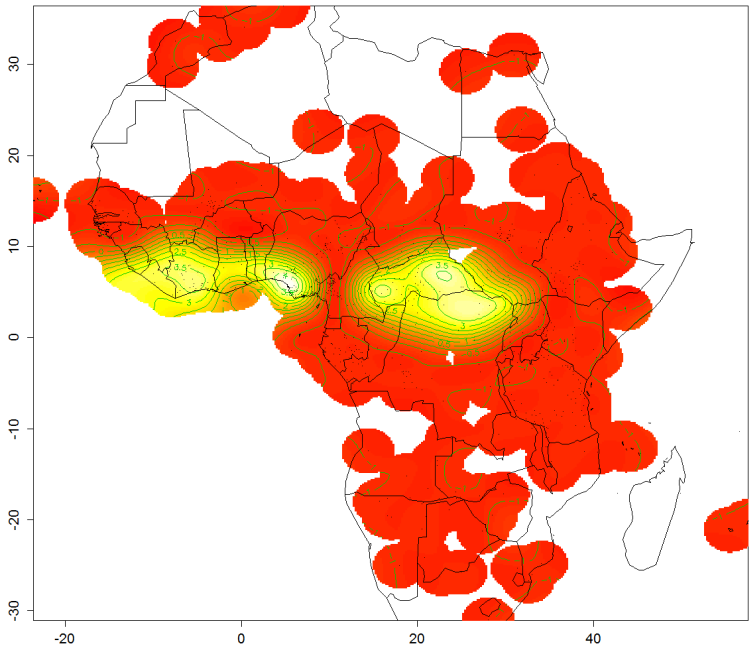


Figure 7. Le tracé de contours produit avec un modèle additif généralisé (k = 17, fonction = scaled-T, edf = 150.7, p < 2e-16, la déviance expliquée = 76.6%, AIC = 1129). Le tracé de contours montre la surface de régression des fréquences lexicales des LV log-transformées en fonction de la combinaison de longitude et latitude en utilisant la méthode de *thin-plate splines* avec le schème des couleurs d'une carte thermique. Les teintes les plus claires correspondent aux fréquences lexicales des LV log-transformées les plus hautes. Les lignes de contours sont des isoglosses qui marquent les écarts de la moyenne en termes d'écart type.

Le GAM visualisé dans la Fig. 7 a une structure moins nuancée que le GAM visualisé dans la Fig. 6 et ressemble plus au simple graphique d'interpolation spatiale dans la Fig. 5. Ainsi, d'un

côté, le GAM visualisé dans la Fig. 7 accentue {p. 203} la discontinuité majeure entre les régions avec une haute fréquence lexicale des LV au niveau du Cameroun et du nord-est du Nigeria, du sorte que le pont des langues à fréquence lexicale des LV basse qui relie dans la Fig. 6 la région centrafricaine et la région côtière par la vallée du Benoue et le plateau d'Adamawa disparait. De l'autre côté, dans la Fig. 7, on ne peut plus discerner la discontinuité au niveau du Ghana et le cluster éventuel dans le sud-est du Mali et le sud-ouest du Burkina-Faso. Le GAM produit en utilisant la fonction *scaled-T* et visualisé dans la Fig. 7 s'avère donc être beaucoup moins informatif que le GAM produit en utilisant la fonction gaussienne et visualisé dans la Fig. 6.

6. Conclusions

Il était bien connu que les occlusives labiales-vélaires (LV), telles que /k͡p/, /g͡b/ et /ŋ͡m/, font partie de l'inventaire phonologique d'un grand nombre de langues du nord de l'Afrique sub-saharienne (NASS), tout en étant très rares ailleurs dans le monde. Nous connaissions également la répartition des langues avec LV dans le NASS (voir, par exemple, Cahill 2008, Clements & Rialland 2008, Maddieson 2011, la base de données Phoible). En même temps, les publications sur les langues individuelles du NASS font parfois état du statut exceptionnel ou marginal des LV dans leurs phonologies, ce qui signalait que la simple énumération des langues avec ou sans LV ne donnait qu'une image très rudimentaire des labiales-vélaires en tant que phénomène aréal. Grâce au développement de la grande base de données lexicale outillée RefLex (Segerer & Flavier 2011-2016), nous avons pu aller au-delà de l'énumeration des inventaires phonologiques qui contiennent des LV et estimer avec précision la fréquence des consonnes labiales-vélaires dans le lexique de 336 langues du NASS. Ces estimations confirment l'hypothèse selon laquelle dans la plupart des langues à consonnes labiales-vélaires, ces consonnes sont peu fréquentes dans le lexique général et encore moins fréquentes dans le lexique de base. Toutefois, nous avons pu identifier deux aires géographiques (dont une composée de deux sous-régions séparées) où la fréquence des consonnes LV est « normale » par rapport à la situation canonique où chaque consonne d'une langue est également fréquente dans son lexique. Ces observations fournissent un ensemble d'indications pour une explication de l'origine et de la diffusion des consonnes labiales-vélaires dans le nord de l'Afrique sub-Saharienne, explication que nous proposerons ultérieurement. {p. 204}

Références :

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. URL: http://www.jstatsoft.org/v12/i06/

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.

CAHILL, Michael Clark. 2008. Why labial-velar stops merge to /gb/. *Phonology* 25. 379–398.

CLEMENTS, Nick & 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.

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.

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

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

PAKENDORF, Brigitte, Koen BOSTOEN & Cesare DE FILIPPO. 2011. Molecular perspectives on the Bantu expansion: A synthesis. *Language Dynamics and Change* 1. 50–88.

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

SEGERER, Guillaume & Sébastien FLAVIER. 2011-2016. *RefLex: Reference Lexicon of Africa*, Version 1.1. Paris, Lyon. URL: http://reflex.cnrs.fr/

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.

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.

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. 2015. *mgcv*. R package version 1.8-6. URL: http://CRAN.R-project.org/package=mgcv

Dmitry IDIATOV & Mark VAN DE VELDE

LLACAN, CNRS, Sorbonne-Paris Cité, INALCO

LLACAN – UMR 8135 du CNRS
7, rue Guy Môquet – BP 8
94801 Villejuif Cedex
France

dmitry.idiatov@cnrs.fr, mark.vandevelde@cnrs.fr