Prédire l'agglutination de l'article en mauricien

Fabiola Henri*, Olivier Bonami**

INTRODUCTION

Les noms hérités du français en mauricien ont souvent une forme analogue à la séquence ARTICLE + NOM du français; une forme qui, sans doute, résulte d'une réinterprétation de frontières de mot. Ce phénomène, présent dans d'autres créoles à base française[1] (CBF), s'avère plus répandu dans les créoles de l'Océan Indien, en particulier en mauricien (Baissac 1880, Baker 1987, Grant 1995b, McWhorter and Parkvall 2002, Stranquist 2003). Comme le note Parkvall (2008), il n'est, en fait, pas étonnant de voir ce type de phénomène de réinterprétation de frontières de mots dans le cadre de contacts de langues qui constituent un terreau naturel pour l'opacité morphologique. Pour les tenants de l'exceptionalisme, (McWhorter 1998, *et. subseq*., Parkvall 2008), l'agglutination de l'article dans les CBF constituerait un argument additionnel en faveur de la créolisation comme transmission interrompue.

Cependant, ce facteur ne saurait être, à lui seul, à l'origine de ces nouvelles formes nominales dans les CBF, étant donné la répartition du phénomène. Cette répartition suscite, en effet, des questions auxquelles certains auteurs ont tenté de répondre. A l'exception de Baissac (1880) qui y voit une influence malgache,

*Université du Kentucky, Département de Linguistique, Courriel : fshe223@uky.edu;
**Université Paris 7, Laboratoire de Linguistique Formelle, Courriel : olivier.bonami@linguist.univ-paris-diderot.fr

***Ce travail a été présenté au troisième colloque *Formal Aspects of Creole Studies* (Lisbonne, 2012), ainsi que dans des conférences à l'ATILF (Nancy), à l'U. Paris Diderot et à l'U. Paris 8. Nous remercions les participants à ces différents événements, et en particulier Alain Kihm et Tonjes Veesntra, pour leurs commentaires et suggestions. Ce travail a bénéficié d'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.

Nous utilisons les abréviations suivantes : LF forme longue, SF forme courte, STF forme forte, WF forme faible, SP défini spécifique, PST passé, FUT future, CONJ conjonction, QU question, IND indéfini

[1] L'agglutination du déterminant est également disponible dans les créoles à base portugaise (Ladhams 2012; Hagemeijer 2009) et sporadiquement, dans quelques autres créoles non-dérivés de langues romanes, comme par exemple, en Sranan et Saramaccan, ou encore dans le pidgin nigérien (Parkvall 2008).

l'agglutination[2] du déterminant en mauricien est réputée dû à l'influence de langues bantoues, contributeurs majeurs de la formation des CBF de l'Isle de France[3] (Baker 1987, Grant 1995b, Strandquist 2003, Veenstra 2011).

Cet article, tout en examinant certaines de ces questions, sera consacré à une question légèrement différente : quels sont les facteurs qui favorisent ou défavorisent l'agglutination du déterminant en mauricien, dans la mesure où tous les noms hérités du français ne sont pas sujets à l'agglutination. Nous identifions certains facteurs qui sont susceptibles de jouer un rôle et nous montrons que ces facteurs favorisent l'agglutination de manière non-catégorique. Nous procédons ensuite à une analyse par régression logistique sur la base de données construites à partir de Carpooran (2011) et d'informations statistiques et phonologiques du lexique du français recueillies dans Lexique 3 (New *et al.* 2007) et dans le *Trésor de la Langue Française*, dans le but de déterminer le degré de pertinence de chaque facteur.

Un deuxième aspect abordé dans cette étude, concerne la prédictibilité de l'agglutination. À ce sujet, une distinction importante doit être faite. Nous appelons *forme agglutinée* un mot dont la forme est la concaténation d'une séquence qui descend d'un déterminant du français et d'une séquence qui descend d'un nom. Nous appelons *forme non-agglutinée* un mot dont la forme descend d'un unique étymon. Certains lexèmes ont une forme unique, qui peut être une forme agglutinée (cf. ZOUR « jour » en mauricien) ou une forme non-agglutinée (cf. LAVI « vie » en mauricien). D'autres lexèmes possèdent deux formes distinctes, dont l'une est agglutinée et l'autre non-agglutinée (cf. DANS « danse » en mauricien, qui a deux formes *dans* et *ladans*). On a donc un contraste à trois termes entre lexèmes non-alternants non-agglutinés, lexèmes non-alternants agglutinés, et lexèmes alternants.
Dans ce travail, nous nous focalisons principalement sur la modélisation du contraste entre les lexèmes donnant lieu à agglutination (qu'ils soient alternants ou non) et lexèmes ne donnant pas lieu à agglutination. L'analyse détaillée des conditions qui président au choix de la forme des lexèmes alternants en fonction de leur contexte d'utilisation ne peut être entreprise ici. Nous soutenons néanmoins l'idée que les formes ne sont pas en variation libre, mais mauricien a innové une fonction à partir de ce contraste de formes ; une fonction morphomique dans la mesure où les contextes d'occurrence des deux formes ne forment pas une classe naturelle.

---

[2] L'agglutination n'est pas à proprement parler un processus morphologique mais un phénomène diachronique de réinterprétation de frontières de mots qui a pour résultat la grammaticalisation de nouveaux mots sur la base d'une combinaison d'au moins deux mots. Nous n'employons pas le terme d'« incorporation nominale » (par exemple, Strandquist (2003), Parkvall (2008), Veenstra 2011) qui porte à confusion dans la mesure où il est utilisé avec un sens technique tout à fait différent en morphologie et en syntaxe.

[3] Le mauricien, le seychellois, le rodriguais et le chagossien.

1. L'ETAT DE L'ART

Le phénomène d'agglutination de l'article du français au substantif qu'il spécifie n'est pas propre au créole. Elle est déjà étudiée par Tappolet (1903) qui note sa présence fréquente dans les variétés informelles ne connaissant guère l'influence de la prescription.

« Les formes d'une langue qui s'écrit sont bien autrement gravées dans la mémoire de ceux qui parlent celles d'un idiome qui ne s'écrit guère. Combien de fois le patois n'hésite-t-il pas entre deux ou trois formes ou façons de dire la où le français académique plus rigide (…) a réduit par élimination l'ancienne variété de formes ! Cet état d'hésitation, propre à tous les idiomes qui ne connaissent pas le correctif de l'écriture, est la cause générale de l'agglutination, qui, en dernier lieu, n'est autre chose qu'une erreur non-corrigée. »
(Tappolet, 1903)

Si elle est rare dans la langue standard, elle n'y est pas inconnue (cf. des formes comme *l-ierre, n-ombril* ou encore *l-uette*) ; elle est particulièrement observée dans la parole des enfants pendant l'acquisition de la langue maternelle (Morin 1982 : 38), aussi bien que dans l'acquisition non supervisée de la langue seconde. Tout laisse donc penser que l'agglutination est le résultat d'un phénomène diachronique général de réinterprétation des frontières de mots, qui est favorisé par les situations de transmission du lexique imparfaites et/ou informelles. Dans ce contexte, la forte prévalence de l'agglutination dans les créoles peut être vue comme une simple conséquence de la situation d'acquisition imparfaite et non-supervisée de la langue lexificatrice qui participe à leur émergence[4].

Ce point de vue n'est guère favorisé par les créolistes, qui voient le plus souvent dans l'agglutination le résultat de simplification (McWhorter et Parkvall 2002) ou d'influences substratiques (Baissac 1880, Baker 1984, Grant 1995b, Strandquist 2003). En particulier, McWhorter et Parkvall (2002) tiennent que le type d'agglutination qu'on retrouve dans les CBF n'est en aucun cas comparable au français dialectal dans la mesure où les CBF exhibent un nombre plus important d'agglutination atypique (NDLR d'agglutination syllabique). L'étude comparative de Grant (1995b) examine 11 CBF sur la base de la liste de Swadesh qui ne contient que 207 éléments et confirme l'observation que le mauricien tend à agglutiner plus l'article au nom que les autres CBF. Cependant, sans données suffisament nombreuses pour permettre une étude statistique, ces affirmations demeurent spéculatives. Wittman et Fournier (1982) attestent d'ailleurs la productivité de l'agglutination du possessif comme *madame* 'dame', *matante* 'tante en français québécois. Ils mentionnent également des données de

[4] On ne peut en outre pas exclure qu'elle ait déjà été présente, dans une certaine proportion, dans les parlers français de l'Ouest du XVIII[e] siècle ayant servi à la lexification des CBF, et constitue donc en partie un phénomène préalable à la créolisation

l'ancien français, où l'agglutination typique précède une seconde étape de déglutination (l'aridelle > la ridelle). En ce qui concerne les CBF, l'agglutination de l'article est beaucoup plus conséquente dans les créoles des Antilles qu'on a pu le croire (Zribi-Hertz et Jean-Louis 2013).

L'agglutination de l'article au substantif constitue, selon un des courant de pensée en créolistique, un des arguments forts en faveur de la thèse de l'exceptionalisme des créoles[5]. La formation des créoles est en effet considérée par plusieurs comme étant le résultat d'une transmission interrompue (McWhorter 1998, *et. subseq*., Parkvall 2008). McWhorter (1998) stipule qu'un créole typique est caractérisé par certaines propriétés structurelles - phonologiques, morphologiques, syntaxiques, sémantiques – qui émergent lors d'un processus de réanalyse dont l'agglutination du déterminant serait un exemple.

> «(…) the exigencies of rapid acquisition combine with lengthy incubation requirements to ensure this situation in creoles. Derivational apparatus, be this affixes or particles, tends to be semantically irregular and, especially in affix form, less salient than free stressed items ; therefore, it tends to be eschewed by creole creators or incorporated in fossilized form.»
> McWhorter (1998 : 798)

> «(…) fossilization, when so rampant as to render a grammar almost completely isolating, is as much a sign of extensive second acquisition as the elimination of affixes. For example, in the French-based plantation creoles, French definite articles and partitive-marking *de* are very often fossilized on nouns, no longer processed as morphemes.»
> McWhorter (2011: 235)

Comme le souligne l'auteur, le contexte particulier de la formation des créoles et la nature de l'input ont façonné ces langues. Il est cependant un peu hâtif d'affirmer que les langues créoles sont des produits exceptionnels et structurellement simples sur cette base. D'abord parce que le déterminant clitique agglutiné n'est pas nécessairement dépourvu de toute fonction : En martiniquais, le *la-* et *le-* préposés au nom marque le défini sémantique et le pluriel respectivement (Zribi-Hertz et Jean-Louis 2013)[6] ; *l-* et *z-* préposés au nom s'oppose en nombre en réunionnais (Caron 2012) ; l'article défini préposé en mauricien semble avoir une fonction morphomique[7] (*cf.* § 1.3). Ensuite, parce que l'agglutination du déterminant, certainement remarquable dans les créoles à

[5] Pour une discussion voir Degraff (2003, 2005).
[6] Zribi-Hertz et Jean-Louis (2013) établissent, en effet, la distribution des formes alternantes comme *plage-laplage* 'plage' en martiniquais et montrent que la forme substantive en *la-* est incompatible avec le déterminant clitique postposé au nom *–la,* comme dans *plage-la*, qui lui s'interprèterait comme un défini pragmatique, i.e., qui ferait référence à une plage visible ou déjà mentionnée dans le contexte.
[7] Une étude approfondie de la distribution des formes alternantes serait nécessaire pour étayer cette hypothèse.

base française, n'est pas généralisable à d'autres créoles à part ceux à base portugaise (Ladhams 2012, Hagemeijer 2009). Si le procédé était caractéristique de la créolisation, on s'attendrait à ce que ce soit plus répandu à travers les créoles. Enfin, la thèse exceptionnaliste n'explique ni la diversité des situations rencontrées parmi les CBF, ni la distribution de l'agglutination, qui n'est pas systématique mais pas non plus aléatoire, comme nous le montrerons plus bas.

Le phénomène serait, pour d'autres, le résultat de changements linguistiques, renforcé par une influence des langues substrats, elles mêmes agglutinantes. La prosthèse de l'article, pour reprendre les termes de Baissac (1880), trouve d'abord son origine dans la langue française du XVe siècle où le substantif s'agglutinait coutumément avec l'article, mais elle est renforcée par le processus d'emprunt de substantifs français en malgache ; processus qu'on retrouve par conséquence également en mauricien étant donné l'origine des esclaves responsable de sa formation. Tout comme son prédécesseur, Baker (1984) propose une explication substratique à la grande proportion de noms agglutinés qu'on retrouve en mauricien. Selon lui, l'origine de l'agglutination est d'influence bantoue. Les nouveaux noms créoles sont ainsi de lexique français mais formés sur le patron morphologique des noms bantous qui portent la marque de leurs classes respectives. Ainsi, pour Baker, mais également Grant (1995b) qui étend l'analyse à d'autres créoles[8], l'abondance des formes agglutinées en mauricien ne serait pas simplement le résultat d'acquisition du français par les esclaves mais serait surtout le résultat d'un processus cognitif étant donné leur langue maternelle. Cette explication de la prévalence originelle de l'agglutination ne saurait cependant être suffisante pour expliquer qu'elle reste fréquente dans des formations qui ont clairement été empruntées au français bien après la créolisation et la disparition de quelque compétence bantoue chez les locuteurs du créole (*lotorout*, *laspirater*, *lelikopter*, etc.).

A côté de cet ingrédient essentiel à la formation de ces nouvelles formes substantivales, Baker mentionne comme facteurs déterminants l'évitement de l'homophonie, la fréquence de collocation de l'article, la monosyllabicité et l'harmonie vocalique. Strandquist (2003) exploite l'idée d'harmonie vocalique dans une étude quantificative basée sur des données lexicales extraites de Baker et Hookoomsing (1987), contenant un total de 518 noms agglutinés des déterminants *la-, le-* et *di(u)-*. *La-*, l'article le plus fréquemment agglutiné de ces données, est favorisé dans la mesure où la voyelle [a] est neutre quant à l'harmonie vocalique et peut apparaître avec n'importe quel nom alors que les voyelles de l'article défini masculin ou du partitif, doivent être en harmonie avec la voyelle interne du nom avec lequel elles apparaissent. Nous reviendrons sur chacun de ces facteurs dans les sections suivantes. Nous ne prendrons, toutefois, aucune position quant à une origine particulière de l'agglutination du déterminant dans cet article. En réalité, nos données et résultats suggèrent qu'il serait plus judicieux de recourir à une explication en termes de facteurs

[8] Voir également Parkvall (2000).

convergents qui impliqueraient les propriétés du lexifieur, les influences substratiques ainsi que des procédés cognitifs.

2. EXPLORATION DE L'AGGLUTINATION DE L'ARTICLE

En mauricien, l'agglutination de l'article prend les formes distinguées dans le Tableau 1[9]. Ces nouvelles formes sont composées de l'agglutination de l'article défini singulier masculin et féminin, de l'article défini neutre singulier et pluriel, de la consonne de liaison du pluriel et de l'indéfini singulier, de l'indéfini pluriel, du partitif et du possessif. Etant donné la nature de ses éléments en français, il aurait été étonnant de ne pas trouver de telles formes agglutinées dans le contexte où elles se sont formées. Ces articles sont, en effet, des clitiques grammaticaux en français (Miller, 1992) : (a) Ils sont sujets à des règles de sandhi comme la liaison et l'élision. (b) Ils sélectionnent un hôte. (c) Ils ne sont pas nécessairement adjacents à leur hôte. Les clitiques grammaticaux sont souvent source d'agglutination (Zwicky 1977). Selon la morphologie diachronique, certaines catégories, tels les pronoms ou les déterminants, peuvent progressivement passer de statut d'éléments syntaxiquement indépendants au statut d'affixes en passant par la catégorie clitique dans les cas d'agglutination atypique ou être simplement agglutinés 'sans motivation morphologique' dans les cas d'agglutination typique (Wittman et Fournier 1982). Le mauricien dispose certainement des deux types dans la mesure où nos données sont composées de formes fixes et de formes alternantes.

TYPE D'ARTICLE	DET. FR	NOM MAU.	SEQUENCE FRAN.
défini	F. SG *la*	latab	la table
	M. SG *le*	ledo	le dos
	SG *l'*	lasiet	l'assiette
	PL. *les*	lizie	les yeux
indéfini	SG. *un*/*une*	nide	une idée
	PL. *des*	dizef	des oeufs
pluriel	DEF. *les* ou IND. *des*	zepis	les/des épices
partitif	*du*	dite	du thé
	de l'	delo	de l'eau
possessif	F. SG *ma*	matant	tante
	M. SG *mon*	monper	prêtre

Tableau 1 Types d'agglutination du déterminant

Comme mentionné donc dans la section introductive, nous nous attacherons ici uniquement à l'examen des facteurs favorisant l'agglutination de l'article défini en mauricien. Nous prenons pour point de départ les propriétés du lexique du

[9] Les formes agglutinées ne sont pas nécessairement les mêmes à travers les CBF.

français contemporain qui sont corrélées avec l'agglutination, parmi lesquelles certaines mentionnées par nos prédécesseurs. Nous rajoutons à cette liste des facteurs traditionnellement établis comme participant à la formation des mots. L'utilisation massive de données quantitatives disponible sur le français et sur le mauricien nous permettra de corroborer ou d'infirmer les observations précédemment formulées.

2.1. Les données prises en compte

Notre étude s'appuiera largement sur une base de données lexicale du mauricien, des données de corpus étant actuellement insuffisantes. Les données diachroniques disponibles (Baker et Fon Sing 2007) sont trop peu nombreuses pour permettre une étude quantitative informative. Les données écrites contemporaines sont par ailleurs difficiles à décrypter pour plusieurs raisons. Premièrement le mauricien a, jusqu'ici, emprunté une graphie plus ou moins française, obscurcissant ainsi certaines distinctions. Certaines formes agglutinées ressemblent souvent à l'écrit à leurs formes d'origine.

(1) *soi mo pou al la messe soi mo pou travay dans la boutik (...)*
soit 1s FUT aller:LF messe soit 1s FUT travailler:SF dans boutique (…)
'Soit j'irai à la messe soit je travaillerai dans la boutique (…)'
(*Verbatim*, Instagram, Clyde Parsoo-Ramen 18/02/2015)

Comme on le constate en (1), l'auteur utilise une graphie et une syntaxe quasi française, même si ces deux noms en mauriciens sont des exemples d'agglutinations fixes. Le mauricien est en phase de standardisation depuis une décennie et n'a été introduit dans le système éducatif, que très récemment, comme langue ancestrale. Il n'est donc pas toujours évident de distinguer de formes héritées des formes empruntées dans les textes disponibles, ou encore comme nous l'avons précédemment mentionné, des traits dialectaux, des traits créoles.

Deuxièmement, l'origine des emprunts de l'anglais est difficile à identifier. Ils peuvent avoir été hérités du français ou directement de l'anglais en mauricien.

Nous nous appuyons donc sur une base de données lexicale de 8800 noms, compilée à partir de la nomenclature du *Diksioner Morisyen* (Carpooran 2011). Nous avons documenté dans entrée la langue source, l'étymon ainsi que la présence de formes alternantes, agglutinées ou non. Comme indiqué dans le tableau 2, nous faisons une distinction entre les noms qui sont hérités du français, et les noms ayant leur origine ultime en français mais créés en mauricien. Ainsi, le nom mauricien *sante* « chanson » descend ultimement du verbe français

chanter, mais pas directement: le verbe *sante* a été hérité du français *chanter*, puis converti en nom en mauricien[10].

EXEMPLE	TRADUCTION	ORIGINE	ÉTYMON	EFFECTIF TOTAL
latab *lestoma*	table estomac	française	*la table* *l'estomac*	7325
sante *koze*	chanson conversation	innovation	*chanter* *causer*	306
zugader *lapang*	joueur bourlet	portugaise malgache	*jogador* *ampango*	1169
			TOTAL	8800

Tableau 2 Exemples de noms mauriciens d'origines variées

Le tableau 3 illustre l'existence de noms n'ayant pas leur origine en français, mais qui commencent cependant par une séquence non-étymologique ressemblant à un article français. Les exemples de ce type ne semblent pas résulter de la réanalyse d'une séquence de mots française, mais plutôt de l'extension à des emprunts à d'autres langues d'une stratégie générale d'adaptation des mots nouveaux au lexique créole.

EXEMPLE	TRADUCTION	ORIGINE
lapang	bourlet	malgache *ampango*
larurut	arrow-root	anglais *arrow-root*
lodians	audience	anglais *audience*
lamar	eau de cuisson du riz	hindi
larak	arak	arabe

Tableau 3 Exemple de noms d'origine non-française agglutinés de l'article défini Fr.

Chaque entrée a été transcrite phonétiquement sur la base de leur graphie. Nous avons ensuite sélectionné un sous-ensemble de 4881 noms issus du français pour lesquelles nous avons des informations pertinentes comme la fréquence et le genre de l'étymon (Lexique 3, New et al. 2007), la phonologie de l'étymon (Flexique, Bonami *et al*. 2014), la fréquence de cooccurrence de l'étymon avec l'article défini (French subtitles corpus, New and Spinelli 2012) et une date de première attestation de étymon dans le *Trésor de la Langue Française*. Enfin, nous avons limité l'étude aux 4760 cas où l'élément agglutiné est l'article défini singulier.

[10] Dans de nombreux cas, où des familles morphologiques parallèles existent en mauricien et en français, il n'est pas possible de décider de manière certaine si un nom dérivé a été hérité du français ou innové en mauricien indépendamment de l'existence du nom français correspondant. Dans ces cas nous avons systématiquement arbitré en faveur de l'héritage.

2.2. Mots fixes

Les mots agglutinés fixes sont ceux qui se rencontrent toujours sous une forme agglutinée. Ils correspondent sans doute à la catégorie des agglutinés typiques dont parlent Wittman et Fournier (1982), c'est à dire, une forme dont la particule agglutinée n'assumerait aucune fonction[11]. C'est le cas des formes comme *latab* 'table', *lisyen* 'chien', *lestoma* 'estomac', *lezel* 'aile', *letour* 'tour'. Ils ne sont syntaxiquement pas contraints et apparaissent avec tous types d'articles.

(2) a. *Bebet-la ena lezel.*
bête-SP avoir:SF aile
'La bête a des ailes.'

b. *So lezel inn kase.*
POSS:3s aile PERF casser:LF
'Son aile a été cassée.'

c. *Enn lezel inn kase.*
IND aile PERF casser:LF
'Une aile a été cassée.'

d. *Lezel-la inn kase.*
aile-SP PERF casser:LF
'L'aile a été cassée.'

2.3. Mots alternants

A côté des mots fixes, le mauricien a innové des mots alternant entre deux formes. Un premier type d'alternants oppose une forme nue à une forme agglutinée du défini.

(3) a. Fr. *baleine* > Mau. *balen/labalen*
b. Fr. *dans*e > Mau. *dans/ ladans*
c. Fr. *jour* > Mau. *zour/lizour*
d. Fr. *accoutrement* > Mau. *akoutreman/lakoutrema*
e. Fr. *coeur* > Mau. *ker/leker*

La distribution de ces alternants ne semble pas correspondre à une fonction unique mais est de toute évidence contrainte.

(4) a. *Sel soupap ki zot ti ena dan lanfer lesklavaz,*
seul soupape que 3p PST avoir:SF dans enfer esclavage,
se lamizik ek ladans aswar apre travay.
PRES musique CONJ danse soir après travail
'La seule soupape qu'ils avaient dans l'enfer de l'esclavage, c'est la musique et la danse le soir après le travail.'

[11] Une analyse alternative serait de considérer que là on retrouve les formes alternantes opposant une forme agglutinée à une forme nue, on a dans les cas de formes fixes une forme syncrétique occupant les deux cases du paradigme.

(D. Palmyre-Florigny, Festival International Kreol 2006)

b. *Bann dans-la ti top*.
PL danse-SP PST top
'Les danses étaient super.'
(Blogspot Abaim, Its party time, 14/08/2010)

c. *Kat zanfan ant laz 7 ek 13 ti fer enn dans hip-hop*
quatre enfant entre age 7 CONJ 13 PST faire:SF IND danse hip-hop
'Quatre enfants entre l'âge de 7 et 13 ans ont fait une danse Hip-Hop.'
(Lalit, Fet travay, Me 2005)

(5) a. *Ki pli lour ? enn liv lapay ou enn liv diplon?*
QU plus lourd IND livre paille ou IND livre plon
'Qu'est qui est plus lourd: une livre de paille ou une livre de plomb?'
(Enigmas on logical thinking, National Inventory Intangible Cultural Heritage 2013)

b. *De roupi, laliv madam*.
deux roupie livre madame
'Deux roupies, la livre madame.'
(Goodman 1971)

(6) a. *Devlop sikomotrisite pou ekrir enn swit mo.*
developer:SF psychomotricité pour écrire:SF IND suite mot
'Développer la psychomotricité pour écrire une suite de mots.'

b. *Mazinn laswit sa zistwar-la.*
imaginer:SF suite DEM histoire-SP
'Imaginez la suite de cette histoire.'
(Ki pase la, Vol. I, Std 2)

Il semblerait dans la plupart des cas la forme nue peut se combiner avec un article indéfini et, plus rarement avec le défini spécifique postposé. Par exemple, si la forme nue du nom *dans* 'danse' en (4b) est possible avec le défini spécifique postposé, les formes nues en (7) n'autorisent pas ce type de spécification. La marque de spécificité dans les cas en (7) n'apparaît qu'avec les formes agglutinées *laliv* 'livre' et *laswit* 'suite'.

(7) a. * *Komye ou dir liv-la* ?
combien FOR:2s dire:SF livre-SP
'Combien vous dites (coute) la livre ?'

b. * *To finn resi ekrir swit-la ?*
WF:2s PERF réussir:SF écrire:SF suite-SP
'As-tu réussi à écrire la suite?'

Lorsqu'elle apparaît sans déterminant, elle a une lecture générique (8a) où elle est ambigüe en termes de nombre (8b).

(8) a. (…) bann rol zom ek fam ki determine par
(…) PL rôle homme CONJ femme KI déterminer:LF par
biolozi ousa sosiete.
biologie ou_soit société

'(…) les rôles des hommes et des femmes qui sont déterminés par la biologie et la société.'

b. Li sanze ar letan, li diferan dan diferan sosiete
3s changer:LF PREP temps, 3s différent dans différent société
'Ça change avec le temps, C'est différent dans différente(s) société(s)'
(Mauritius Local Government Gender Action Plan)

A l'inverse, les formes agglutinées sont généralement incompatibles avec l'indéfini.

(9) a. * *Donn mwa enn laliv.*
donner:SFSTF:1s IND LA-livre
'Donnez-moi une livre.'

b. ?? *Nou'nn dans enn ladans dan mariaz.*
1p'PERF danse:SF IND danse dans mariage
'Nous avons dansé une danse au mariage.'

c. * *Nou dan enn lasosiete bien koronpi.*
1p dans IND société bien corrompu
'Nous sommes dans une société bien corrompue.'

Mais elles le sont avec le défini spécifique postposé. Comme on le constate en (10), les formes en *la-* sont grammaticales en présence du défini postposé, périphérique[12] en mauricien. Les phrases en (10) ne sont licites que dans un contexte où le nom défini a déjà été mentionné précédemment. Par exemple (10b) est possible en tant que question écho dans un contexte où le locuteur souhaite confirmer le prix donné par le poissonnier.

(10) a. *Komie ou dir laliv-la ?*
combien FOR:2s dire:SF livre-SP
'Combien vous dites coûte la livre ?'

b. *Lasosiete-la inn koronpi net.*
société-SP PERF corrompu totalement
'La société est totalement corrompue.' (société spécifique)

c. *Kouma to dir ladans-la ti ete ?*
comment WF:2s dire:SF danse-SP PST COP:LF
'Comment tu dis que la danse était ?' (danse spécifique)

C'est également ce qu'affirment Zribi-Hertz et Jean-Louis (2013) des formes agglutinées en *la-* et *le-* du martiniquais. L'analyse qu'ils proposent est que les formes agglutinées expriment un défini sémantique dénotant des individus à la manière noms propres (défini congruent). L'article postposé au nom, enclitique en martiniquais, est lui, analysé comme un défini pragmatique (défini incongruent) ; d'où leur incompatibilité. Cette proposition semble compatible avec les données du mauricien. Cependant, les formes alternantes du mauricien

[12] L'article postposé est en mauricien un *edge clitic* traduit ici par clitique périphérique. Il peut être séparé du nom tête par une phrase relative.

ont un comportement moins régulier que ceux du martiniquais : certaines formes agglutinées autorisent une combinaison avec le défini postposé (11).

(11) *Laswit-la pa bon ditou.*
suite-SP NEG bon du_tout
'La suite n'est pas bonne du tout.'

La distribution des mots alternants n'étant pas particulièrement notre objet d'étude dans cet article, nous remettons à plus tard une recherche plus détaillée du phénomène. Cependant, quelques observations sont de mise. D'abord, comme nous l'avons évoqué précédemment, la distribution de l'alternance de forme est contrainte mais ne correspond pas une distribution fonctionnellement cohérente. Ce comportement suggère que la distribution des mots alternants entre forme agglutinée du défini et forme nue, est morphomique au sens d'Aronoff (1994). Ensuite, comme le fait remarquer Zribi-Hertz et Jean-Louis (2013), la simplification ne pourrait expliquer un tel développement, puisque là où le français utilise une même forme pour exprimer les deux types de défini, sémantique et pragmatique, certains CBF ont élaboré un paradigme morphologique plus complexe pour les distinguer ; sans oublier le fait que les CBF au contraire du français autorisent des SN nus. La grammaire du français est sans aucun doute à l'origine de cette évolution, étant donné l'ambiguïté des syntagmes nominaux définis en français et les propriétés morpho-phonologiques de l'article défini. Une hypothèse similaire peut être propose pour l'article postposé. Elle ne contredit aucunement une influence convergente du substrat. Cette évolution suppose d'autre part que les locuteurs du créole ont nécessairement identifié ces différentes fonctions qui auraient été attribuées aux diverses formes perçues. Cependant, nous ne sommes pas en mesure de déterminer la date d'apparition des différentes formes dans le lexique du mauricien compte tenu de l'insuffisance des données diachroniques. Il est donc impossible de savoir si les formes alternantes, opposant une forme agglutinée du défini à une forme nue, sont apparues simultanément. Mais un simple aperçu de l'histoire de certains noms montre que l'alternance est possible aussi bien avec des noms présents dans la langue dès l'époque de la créolisation (*koloni, dans*) qu'avec des noms indiscutablement empruntés au français (*aeropor, antet*).

FORME NUE	FORME AGGLUTINEE	1E ATTESTATION	TRAD.
koloni	lakoloni	1308	colonie
aeropor	laeropor	1922	aéroport
antet	lantet	1838	entête
dans	ladans	1172	danse

Tableau 4 Noms héritées et leur date de première attestation en français.

Une étude de corpus extensive serait bien sûr nécessaire pour confirmer ces hypothèses.

Notons, avant de conclure cette section, que notre base de données comporte d'autres alternances de types variés. Nous pouvons distinguer des alternants en variation libre, par exemple *nide-lide* 'idée'. Nous disposons de cas de glissement de sens comme *listwar* 'histoire' (d'un point de vue collectif) vs *zistwar* 'histoire' (d'un point de vue individuel). Enfin, les rares cas d'agglutination du possessif expriment comme pour les agglutinés du défini, un défini sémantique, ex. *monper* 'père' (certains membres du clergé) vs à un titre *per* 'père'.

Nous limitons donc cette étude à la prédictibilité de l'agglutination de l'article défini au nom en mauricien, qu'elles soient fixes ou alternantes.

3. FACTEURS PRIS EN COMPTE

Les recherches précédentes sur l'agglutination de l'article en mauricien ont suggéré que l'agglutination était sensible à un certain nombre de facteurs. Dans cette section nous reprenons ces facteurs un par un et utilisons des tests statistiques pour évaluer leur pertinence.

3.1. Nombre de syllabes

En mauricien comme dans beaucoup d'autres langues, la longueur moyenne d'un mot est légèrement supérieure à deux syllabes. Les mots très longs sont défavorisés, et les mots d'une seule syllabe tendent à avoir des propriétés prosodiques suboptimales. Pour ces raisons, on s'attend à ce que, face à la possibilité d'agglutiner ou non l'article, la prévalence de l'agglutination diminue avec la longueur de l'étymon, puisque (pour les étymons commençant par une consonne) l'ajout de l'article rend la longueur du mot optimale pour les monosyllabes et la rend suboptimale pour les polysyllabes.

La Figure 1 indique comment la prévalence de l'agglutination varie en fonction de la longueur en syllabes de l'étymon français (en valeur absolue à gauche, en proportion à droite). Une simple lecture graphique suggère que la prévalence de l'agglutination varie notablement suivant que l'étymon est monosyllabique ou polysyllabique. Cet effet est significatif (χ^2=160,76 ; df=1 ; *p*-valeur < 0,0001). En revanche, on ne remarque pas d'effet monotone de la longueur du mot. Au contraire, une fois que le mot fait plus d'une syllabe, la prévalence de l'agglutination décroit plutôt avec la longueur. Cet effet est tout juste significatif (χ^2=9,92 ; df=4 ; *p*-valeur < 0,042), et une étude de plus grande ampleur serait nécessaire pour déterminer s'il est réellement pertinent.

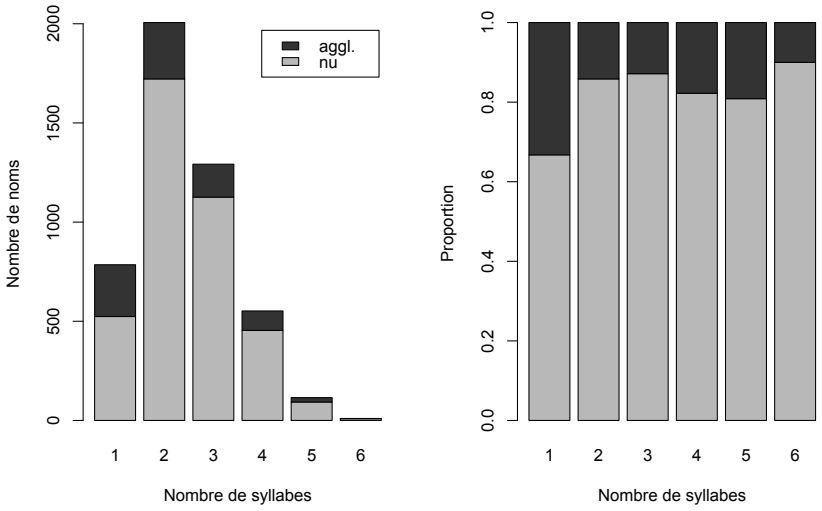

Figure 1 Relation entre prévalence de l'agglutination nombre de syllabes de l'étymon

3.2. Le segment initial de l'étymon

Les langues préfèrent en général les mots à initiale consonantique aux mots à initiale vocalique. Il se trouve que quand un mot français est à initiale vocalique, l'agglutination de l'article en fait un mot à initiale consonantique; alors que quand il est à initiale consonantique, l'agglutination ne produit pas de changement de classe du segment initial, qui reste consonantique. Toutes choses égales par ailleurs, on s'attend donc à ce que les étymons à initiale vocalique préfèrent l'agglutination[13].

[13] Un relecteur anonyme suggère que le mauricien comporte peu de mots à initiales vocaliques, ce qu'on pourrait interpréter comme un facteur favorisant l'agglutination de l'article. Les données disponibles ne confirment pas cette impression. Examinons la situation des verbes, pour lesquelles la question de l'agglutination de l'article ne se pose pas. Carpooran (2011) contient 504 verbes à initiale vocalique sur 2079, soit 24,2%. En comparaison, la base de données Flexique (Bonami et al., 2013) 24,6% de verbes à initiales vocaliques, soit presque exactement le même nombre. Nous concluons que rien n'indique une préférence du mauricien pour les mots à initiale vocalique supérieure à celle qui prévaut pour la langue lexificatrice.

Les données quantitatives confirment un rôle significatif de la classe du segment initial de l'étymon (χ^2=973,88; df=1 ; *p*-valeur < 0,0001), illustré dans la figure 2.

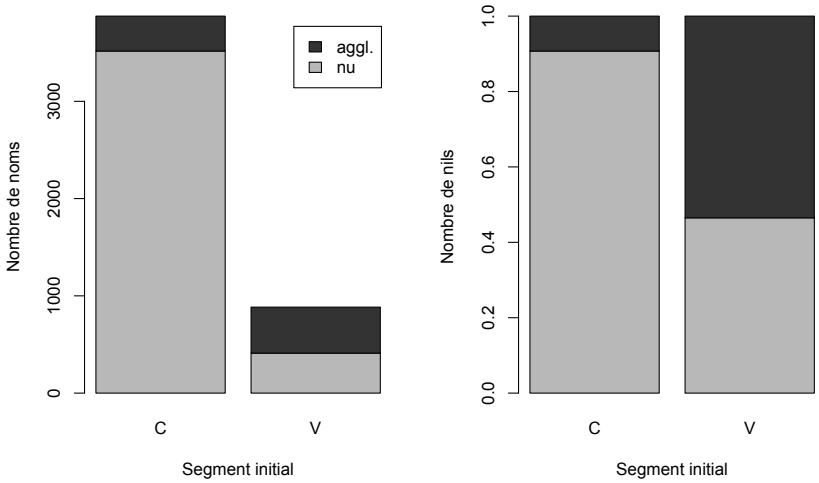

Figure 2 Relation entre la prévalence de l'agglutination et le type du segment initial de l'étymon

3.3. Les effets de dissimilation

Les phénomènes de création de lexème sont souvent soumis à des effets de dissimilation (Plénat 2000) : toutes choses égales par ailleurs, la création d'un nouveau lexème est défavorisée si celle-ci conduit à placer côte à côte deux segments identiques ou proches. Ceci est souvent la cause soit d'une troncation (*milanais*>*milanissime* et non *milanaisissime*), soit du choix d'un autre procédé de formation (*camion*>*camionneur* et non *camionnier*). Si on rapproche la l'intégration d'un lexème en créole d'un procédé de création de lexème, on pourrait s'attendre à rencontrer des effets de ce type; spécifiquement, il se pourrait que l'agglutination soit défavorisée avec les bases commençant par /l/, voire /la/ ou /le/.

Nos données ne confirment pas l'existence d'un effet de dissimilation. En réalité, comme le montre la Figure 3, il y a un léger effet contraire, mais à peine significatif (χ^2 = 4,3965 ; df=1 ; *p*-valeur ≈ 0.036) : l'agglutination est légèrement plus fréquence avant /l/ qu'avant les autres consonnes.

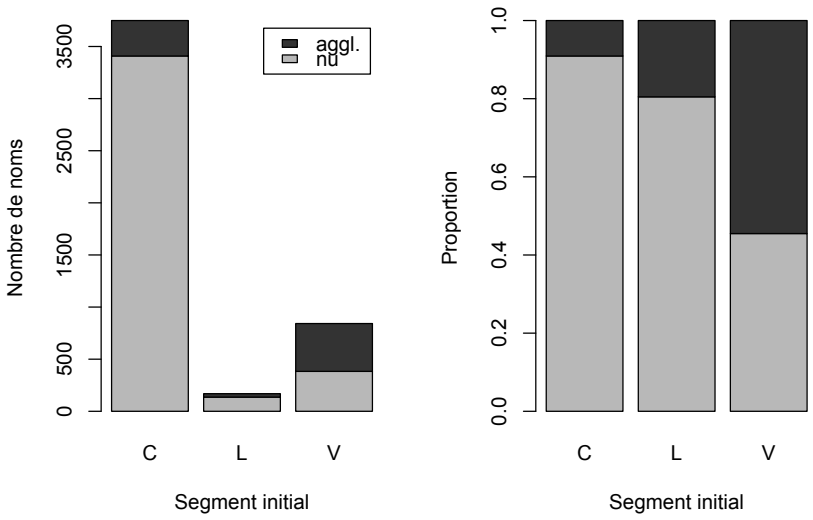

Figure 3 Effets de dissimilation

3.4. Harmonie vocalique

Strandquist (2003) soutient que l'agglutination est favorisée quand la voyelle de l'article est en harmonie avec le nom. En cas de disharmonie en français, soit l'agglutination n'a pas lieu, soit l'harmonie est innovée par le créole. Cette harmonisation ne concerne que les voyelles [e] vs [i], le [a] étant neutre (§ 1).

(12)	a.	*ləky* 'le cul'	>	*liki* 'cul'
	b.	*ləʃjɛ̃* 'le chien'	>	*liʂjɛ̃* 'chien'
	c.	*ləʃu* 'le chou'	>	*lisu* (*var.* lesu) 'chou'
	d.	*ləli* 'le lit'	>	*lili* 'lit'

L'idée que l''harmonie vocalique influe sur l'agglutination est intrinsèquement liée à celle d'une influence des langues bantoues sur la créolisation, qui elles-mêmes sont sujettes à un tel processus dans le domaine nominal.

L'examen de nos données fait cependant fortement douter de l'importance de ce facteur. D'une part, son influence n'est certainement pas catégorique, puisqu'on trouve des exemples comme ceux de (13). D'autre part, même s'il jouait un rôle,

celui-ci ne pourrait être qu'épiphénoménal, dans la mesure où seuls 1,3% de nos noms (60 items) ont une première syllabe candidate à l'harmonie vocalique avec l'article.

(13) a. *letour* ~ tour 'tour'
b. *ledwa* 'doigt'
c. *lekouyon* ~ *kouyon* 'couillon'
d. *lerim* 'rhume'
e. *lelou* 'loup'
f. *leziz* 'juge'

3.5. Le genre de l'étymon

Dans la mesure où l'article prend une forme différente en français selon le genre du nom, on s'attend à la possibilité que la prévalence de l'agglutination varie selon le genre de l'étymon français d'un nom mauricien. Nos données confirment que c'est le cas. Dans la figure 4, les noms français sont classés en trois catégories, selon qu'il s'agit de noms toujours masculins (comme *bateau* ou *chanteur*), toujours féminins (comme *table* ou *chanteuse*), ou ambigus entre masculin et féminin (comme *secrétaire* ou *pianiste*). Nous ne faisons pas de différence entre les situations d'homonymie, de polysémie, ou de neutralisation de l'opposition de genre. La figure indique clairement que les noms masculins ou ambigus donnent moins souvent lieu à agglutination que les noms féminins. Cette différence est significative (χ^2=206,75 ; df=1 ; *p*-valeur < 0,0001).

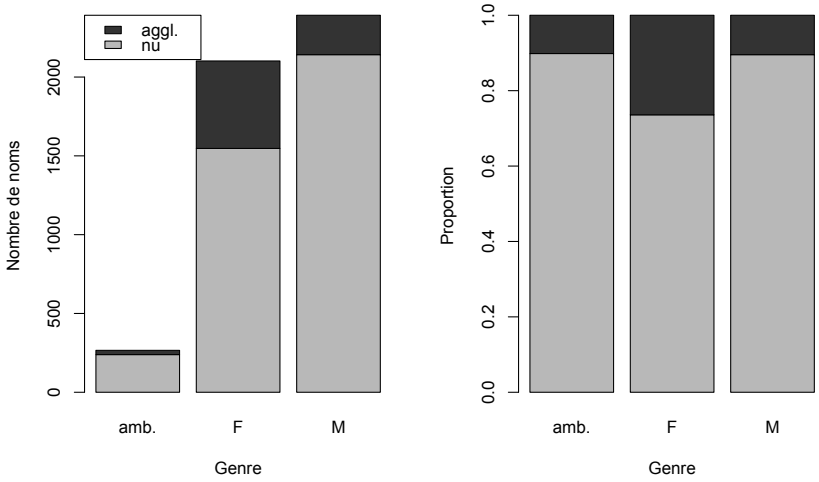

Figure 4 Relation entre prévalence de l'agglutination et genre de l'étymon

3.6. Homophonie

Un autre facteur proposé comme favorisant l'agglutination est l'homophonie (Baker 1987). L'agglutination serait favorisée s'il existe dans le lexique un deuxième mot de forme identique, qu'il soit de même catégorie ou non. Ceci expliquerait pourquoi *lamor* plutôt que *mor* est utilisé comme nom, *mor* étant spécialisé comme forme d'adjectif. Ce facteur est au plus une tendance, puisqu'il existe des cas de formes alternantes où la forme nue correspond dans beaucoup de cas à une forme du verbe (dans$_V$ 'danser' - ladans$_N$ 'danse', danse$_V$ 'danser' - danse$_N$ 'dancing').

Notre base de données ne nous permet pas d'étudier les effets de l'homophonie en général (dans la mesure où nous ne disposons pas d'un lexique complet sous forme électronique), mais nous pouvons le faire dans le cas particulier de l'homophonie avec un verbe. Dans ce cas particulier, on ne rencontre pas l'effet attendu, mais un effet contraire statistiquement significatif ($\chi^2 = 7.0534$; df=1 ; *p*-valeur $\approx$ 0.008). Ce résultat partiel nous pousse à être sceptiques quant à l'existence d'un effet favorable de l'homophonie dans le cas général.

Figure 5 Homonymie

3.7. L'âge de l'étymon

Comme nous l'avons signalé plus haut, l'agglutination de l'article se rencontre aussi bien avec des noms dont tout laisse penser qu'ils appartiennent au lexique créole depuis la phase de créolisation qu'avec des noms qui soit ont été créés en créole, soit ont été emprunté au français ou à une autre langue à une date plus récente. Ce fait étant établi, il reste qu'on s'attend à la possibilité que la prévalence de l'agglutination ait varié dans le temps, l'influence du français n'étant de toute évidence pas de la même nature ni de la même intensité suivant les époques.

Pour étudier l'influence de l'époque d'intégration au lexique créole, l'idéal serait de disposer d'informations philologiques détaillées sur celui-ci. Or, comme on l'a déjà noté, le corpus existant de textes créoles est d'une taille beaucoup trop limitée pour permettre une telle étude. Faute de mieux, le facteur que nous prendrons en compte est donc l'âge de l'étymon français, tel qu'il peut être estimé à partir de la date de première attestation donnée par le *Trésor de la langue française*. Bien que la relation soit très indirecte, il y a de toute évidence une corrélation entre la date de première attestation et la date de première utilisation effective d'un nom, mais aussi entre la date de première attestation en français et la date de première utilisation en créole: un nom d'étymologie française ne peut pas avoir été intégré au créole avant d'exister en français.

La figure 6 illustre l'existence dans nos données d'une corrélation entre l'âge de l'étymon français (groupé par siècle) et la proportion d'agglutination, dont la significativité est confirmée par une régression linéaire (test du rapport de vraisemblance: χ^2=163,20 ; df=1 ; *p*-valeur < 0,0001). Il est nécessaire d'interpréter la figure avec précaution. D'une part, comme l'indique la partie gauche, il y a très peu d'étymons attestés avant le XI^{e} siècle, ce qui fait que l'estimation des proportions pour cette période est incertaine. D'autre part, bien qu'une corrélation soit clairement présente y compris pour le sous-ensemble des noms attestés avant le XVII^{e} siècle, rien ne laisse penser que celle-ci nous informe sur un effet de l'époque d'introduction du nom dans le créole, puisque tous ces noms existaient avant la créolisation et sont donc candidats à faire partie du français auquel les premiers locuteurs créoles ont été exposés. Il est donc plus que probable que, au moins pour les noms ayant une attestation ancienne, une autre cause non-identifiée qui corrèle avec l'âge de l'étymon explique les observations ; par exemple, il est plausible qu'il existe une corrélation entre la date de première attestation et la fréquence de l'étymon dans le français des colons (les mots plus fréquents ayant plus de chances d'être attestés anciennement). À l'inverse, pour les étymons les plus récents, il est plausible (mais en aucun cas certain) que l'âge de l'étymon nous informe en partie sur l'époque d'introduction du nom dans le créole.

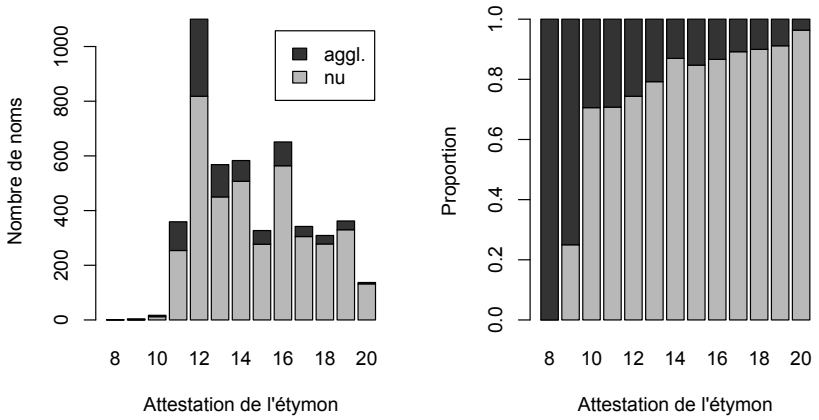

Figure 6 Relation entre prévalence de l'agglutination et date d'attestation de l'étymon (par siècle)

3.8. La fréquence brute de l'étymon

Il est concevable que la fréquence de l'étymon d'un nom mauricien ait une influence sur la probabilité qu'il donne lieu à agglutination : toutes choses égales par ailleurs, plus un nom est fréquent, plus il est probable que les locuteurs aient eu l'occasion de le rencontrer en combinaison avec des déterminants variés, ce qui rend plus improbable une erreur de segmentation.
Pour évaluer l'existence d'un tel effet, et en l'absence de données en quantité suffisante sur le français ordinaire à des époques variées, nous devons nous contenter d'une estimation en français contemporain, et faire l'hypothèse que celle-ci n'est pas trop éloignée de celle qui avait court au moment où chaque nom a été introduit dans le créole. À cette fin nous utilisons le *French Subtitles Corpus* (New et Spinelli, 2013). La figure 7 illustre graphiquement le fait que la fréquence de l'étymon favorise l'agglutination[14], ce qui est confirmé par une régression logistique (test du rapport de vraisemblance : χ^2=274,55 ; df=1 ; *p*-valeur < 0,0001). Il est notable cependant que l'effet va dans la direction inverse de celle qui était attendue.

[14] Pour faciliter l'illustration graphique, les noms ont été rangés par fréquence de l'étymon puis groupés en 50 quantiles; les proportions d'agglutination sont données pour ces quantiles. La courbe est une courbe de régression locale sur ces quantiles.

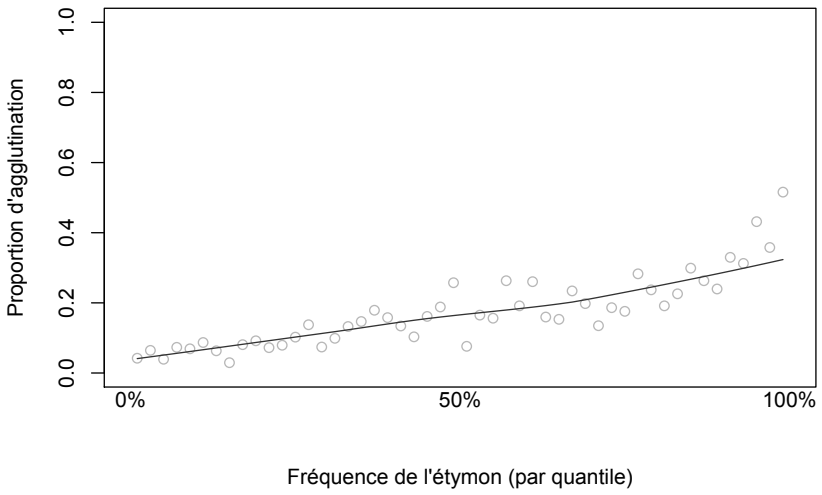

Figure 7 : relation entre fréquence de l'étymon et prévalence de l'agglutination

3.9. Fréquence de collocation avec l'article

Un nom toujours accompagné de la même la forme d'article est plus susceptible d'être agglutiné qu'un nom qui alterne entre divers déterminants. Cette probabilité est nettement plus élevée si la fréquence de la collocation des deux est élevée dans l'usage ; ce qui expliquerait l'agglutination des séquences comme 'la lune'.

Pour évaluer la réalité d'un tel effet, nous examinons le rôle de la fréquence de collocation entre l'étymon et l'article défini singulier, toujours sur la base du *French Subtitles Corpus*. La figure 8 illustre l'existence d'une corrélation positive entre fréquence de collocation et proportion d'agglutination[15], corrélation qui est confirmée par une régression logistique (test du rapport de vraisemblance: χ^2=431,74 ; df=1 ; *p*-valeur < 0,0001).

[15] Pour faciliter l'illustration graphique, les noms ont été rangés par fréquence de l'étymon puis groupés en 50 quantiles; les proportions d'agglutination sont données pour ces quantiles. La courbe est une courbe de régression locale sur ces quantiles.

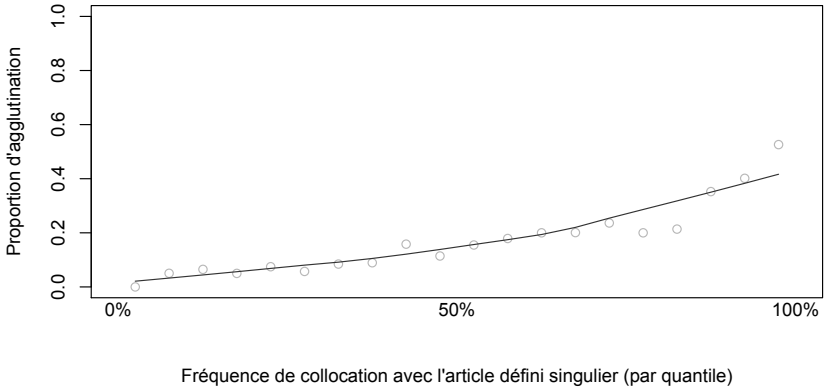

Figure 8 : Relation entre fréquence de collocation et prévalence de l'agglutination

3.10. Bilan

Dans cette partie nous avons confirmé quantitativement que six facteurs favorisent l'agglutination:
(i) Les noms dont l'étymon français est monosyllabique agglutinent plus souvent l'article que les noms dont l'étymon français est polysyllabique.
(ii) Les noms dont l'étymon est à initiale vocalique agglutinent plus souvent l'article que les noms dont l'étymon est à initiale consonantique.
(iii) Les noms dont l'étymon est féminin agglutinent plus souvent l'article que les noms dont l'étymon est féminin ou ambigu quant au genre.
(iv) Plus l'étymon est attesté depuis longtemps, plus il est probable que le nom agglutine l'article.
(v) Plus l'étymon est fréquent, plus il est probable qu'il agglutine l'article.
(vi) Plus l'étymon est souvent accompagné de l'article défini singulier, plus il est probable que le nom ait agglutiné cet article.

En revanche nous n'avons pu confirmer ni effet de dissimilation, ni effet du à l'harmonie vocalique, ni effet de l'homonymie de la forme non-agglutinée avec un verbe.

De toute évidence, d'autres facteurs, notamment syntaxiques ou sémantiques, pourraient être pertinents. Nous nous sommes limités ici aux facteurs qui sont facilement opérationnalisables à grande échelle à partir des ressources disponibles sur le français et la mauricien.

4. MODELISATION STATISTIQUE

Les observations présentées dans la section précédente montrent que l'adoption d'une forme agglutinée ou non-agglutinée pour un nom mauricien est conditionnée par un certain nombre de facteurs non-déterministes, qui relèvent à la fois de la phonologie, de la morphosyntaxe, de la structure et de l'histoire du lexique, et de l'usage. L'étude des facteurs pris un par un est cependant insatisfaisante, dans la mesure où les facteurs sont susceptibles d'être corrélés entre eux ; de ce fait, il est concevable, par exemple, que le rôle apparemment joué par le genre soit illusoire, et que l'impression que le genre joue un rôle s'explique par le fait que les noms féminins sont plus souvent monosyllabiques que les noms masculins (20% vs. 13%). On souhaiterait donc maintenant savoir comment les facteurs dont nous avons décrit le comportement individuel agissent conjointement et quel est le rôle de chacun une fois pris en compte le rôle de tous les autres. À cette fin, nous en œuvre une modélisation statistique multivariée en utilisant la régression logistique. La régression logistique est un outil statistique qui permet de modéliser le comportement d'une variable binaire en fonction de plusieurs variables prédictrices[16]. Dans le cas qui nous intéresse, la variable binaire à prédire est le choix entre agglutination et non-agglutination, et les variables prédictrices sont l'ensemble des facteurs que nous avons décrits dans la section précédente.

Plus précisément, la régression logistique vise à évaluer, pour chaque nom mauricien d'origine française, le poids optimal qui doit être attribué à chaque variable prédictrice pour évaluer aussi correctement que possible la probabilité que le nom agglutine l'article défini. Si Y est la variable à prédire, $\vec{X}$ le vecteur des valeurs des variables prédictrices et $\vec{\beta}$ le vecteur des coefficients estimés par la régression pour ces variables, alors le calcul de la probabilité se fait grâce à la formule suivante[17] :

$$P(\text{agglutination}|\vec{X}) \quad \frac{e^{\alpha - \vec{\beta}\vec{X}}}{1 + e^{\alpha - \vec{\beta}\vec{X}}}$$

Dans le cas présent, et comme l'illustre le tableau 1, il s'avère que toutes les variables précédemment considérées jouent un rôle significatif dans la prédiction

[16] Voir Baayen (2008, chap. 6) pour une introduction à l'utilisation de la régression logistique pour l'étude de données lexicales.

[17] L'*intercept* α est une constante qui permet de calibrer la probabilité correcte dans le cas où toutes les variables prennent leur valeur non-marquée; dans le cas présent, les variables scalaires ont été normalisées et centrées sur la moyenne, et les valeurs considérées comme non-marquées pour les variables nominales sont celles qui ne sont pas mentionnées dans le tableau 6: polysyllabique, initiale consonantique, et masculin.

de l'agglutination: pour chaque variable, le test de Wald produit une *p*-valeur inférieure à 0,0001[18].

	Coefficient	Erreur standard	*z*-valeur	*p*-valeur
Intercept	-4,4573	0,1422	-31,35	<0,0001
monosyllabique	1,9814	0,1382	14,34	<0,0001
initiale vocalique	3,8591	0,1402	27,54	<0,0001
log. fréquence	0,6507	0,0631	10,31	<0,0001
freq. collocation	0,8935	0,0523	17,09	<0,0001
féminin	1,2750	0,1125	11,33	<0,0001
siècle	-0,3156	0,0628	-5,03	<0,0001

Tableau 6 : Régression statistique

Le rôle de chaque variable prédictrice ayant été confirmé, il convient de s'interroger quant aux raisons pour lesquelles certaines variables ont l'effet qu'elles ont. Trois variables ont un effet surprenant: la fréquence brute, le genre, et l'âge de l'étymon.

En ce qui concerne la fréquence brute, nous n'avons à l'heure actuelle pas d'explication à proposer pour le résultat observé. En ce qui concerne les deux autres prédicteurs, nous proposons l'hypothèse suivante: l'âge de l'étymon pourrait être pertinent au sens où les mots qui appartiennent au créole depuis sa formation se comportent différemment des mots empruntés au français par des locuteurs natifs du créole. Pour en juger, nous avons procédé à des régressions logistiques séparées pour les mots à étymon ancien (avant 1800) et récents (après 1800). Le tableau 7 résume les résultats.

	ETYMONS ANCIENS		ETYMONS RECENTS	
PREDICTEURS	coeff.	*p*-valeurs	coeff.	*p*-valeurs
monosyllabique	2,0083	< 0,0001	0,9618	0,2313
initiale_vocalique	3,8709	< 0,0001	3,5815	< 0,0001
fréquence	0,6230	< 0,0001	0,6997	0,0030
fréquence_colloc_def.sg	0,8836	< 0,0001	0,9503	< 0,0001

[18] Rappelons que cela signifie que, pour chaque variable, si cette variable n'avait aucun pouvoir prédictif, la probabilité d'observer les données que nous observons serait inférieure à 1 sur 10000.
Le modèle obtenu est à la fois fortement prédictif (avec un indice de corrélation par rang de Summer $D_{x,y}$=0,846) et peu surajusté (sur 200 rééchantillonages, $D_{x,y}$ est en moyenne à 0,845).

féminin	1,3261	< 0,0001	-0,3621	0,4583
siècle_attestation	-0,2604	< 0,0001	-0,3835	0,2311
	$N = 4261$		$N = 499$	
	$Dx,y = 0,839$		$Dx,y = 0,890$	

Tableau 7 Régressions logistiques pour les noms anciens et récents

Nous constatons que, si tous les facteurs jouent un rôle ancien pour les mots apparus avant la créolisation, pour les mots plus récents, la monosyllabicité, le genre et l'âge semblent avoir cessé de jouer un rôle. L'absence d'un rôle de la monosyllabicité pour les noms récents s'explique naturellement par le trop petit nombre de données (seulement 38 noms monosyllabiques après 1800). De a même manière, il n'est pas étonnant que le siècle d'attestation n'ait pas d'effet significatif pour les noms récents, dans la mesure où l'effet est de faible ampleur sur l'ensemble des données et où les noms récents constituent un sous-ensemble de moins de 10%.

Ce qui reste à expliquer est donc le rôle différent joué par le genre de l'étymon pour les noms anciens et récents. Nous interprétons cette différence comme soutenant l'hypothèse d'une influence du substrat bantou (Baker, 1984; Grant, 1995b). Pour rappel, l'hypothèse de l'influence substratique suppose que l'article français a été traité par les créateurs du créole comme analogue à un marqueur de classe nominale bantoue. Or l'article féminin ressemble plus à un marqueur de classe bantoue que le masculin ; Le féminin a une voyelle pleine, le masculin n'a qu'un schwa qui est souvent non-réalisé. La non-réalisation du schwa rend le *l* du masculin peu saillant et difficilement distinguable d'un *l* prévocalique qui ne marque pas le genre. Si une influence bantoue a eu cours au moment de la créolisation, on s'attend donc à ce qu'elle ait conduit à favoriser l'agglutination surtout pour les noms à étymon féminin. En revanche, rapidement après la période de créolisation, l'adoption généralisée du créole comme langue maternelle par les esclaves a entraîné la disparition de l'influence bantoue, et donc la disparition de la préférence pour l'agglutination au féminin.

5. CONCLUSIONS

Dans cet article nous avons montré que l'agglutination de l'article a existé pendant toute l'histoire du mauricien, et continue à avoir lieu pour les nouveaux emprunts. Nous avons identifié deux types de noms: des noms à forme fixe, agglutinée ou non, et des noms à forme variable, avec un conditionnement du choix de la forme qui reste à examiner en détail. Des causes variées contribuent à expliquer la distribution de l'agglutination, sans qu'aucune combinaison de causes joue un rôle catégorique. Nous avons montré que les causes mentionnées ici ont évolué au fil de l'histoire de la langue. Nous avons proposé un scénario

compatible avec les données quantitatives et historiques sur les raisons de cette évolution.

Il est bien entendu que l'interprétation de ces résultats est limitée par la quantité et la qualité des données disponibles. Nous ne disposons actuellement de données ni sur le français dialectal des colons ni sur les états de langue intermédiaires entre la créolisation et la période contemporaine. Si notre travail est empiriquement plus solide que les précédents en termes de données, il est cependant restreint quant à la taille du corpus sur le mauricien contemporain et bénéficierait certainement d'un lexique français étendu annoté sémantiquement. Il serait donc crucial d'avoir accès à de très grands corpus contenant par exemple, des données de collocation sur des noms rares ; ce qui nous permettrait de mieux comprendre la corrélation entre fréquence et agglutination. Étant donné ces limitations, il est étonnant qu'on aboutisse à une qualité de prédiction non-triviale. Néanmoins, l'analyse statistique de la constitution du lexique est payante même en l'absence de données détaillées.

BIBLIOGRAPHIE

Allen, R. B., 2008, 'The constant demand of the French: The Mascarene slave trade and the worlds of the Indian Ocean and Atlantic during the 18th and 19th centuries.' *The Journal of African History* 49, p. 43–72.

Aronoff, M., 1994, *Morphology by Itself*. Boston: MIT Press.

Baissac, C., 1880, *Études sur le patois créole mauricien*. Nancy.

Baayen, R. H. 2008, *Analyzing linguistic data. A practical introduction do statistics using R*. Cambridge: Cambridge University Press.

Baker, P., 1984, 'Agglutinated French articles in Creole French: Their evolutionary significance'. *Te Reo*.

---------- 2007, 'Elements for a sociolinguistic history of Mauritian Creole (to 1968)'. In P. Baker and G. Fon-Sing (eds.), The making of Mauritian Creole. London : Battlebridge Publications, p. 307–340.

Bonami, O., Boyé, G., et Henri, F., 2011, 'Measuring inflectional complexity : French and Mauritian'. Papier presenté au Workshop *Quantitative Measures in Morphology and Morphological Development*. San Diego : University of California.

Bonami, O., Gauthier C., et Plancq, C., 2014, 'Construction d'un lexique flexionnel phonétisé libre du français'. *Actes du quatrième Congrès Mondial de Linguistique Française*, p. 2583-2596.

Bonami, O., Henri, F., et Luís, A. R., 2012, 'Tracing the origins of inflection in Creoles : a quantitative analysis'. Paper presented at the 9th Creolistics Workshop, Aarhus, Denmark.

Carpooran, A., 2011, *Diksioner Morisien*. Sainte Croix (Mauritius): Koleksion Text Kreol, 2nd Edition.

Goodman, M. F., Nussbaum V L., et Overbey D., 1971, *Mauritian Creole: an introduction*. Washington: Center for Applied Linguistics.

Grant, A P., 1995a, *Agglutinated nominals in Creole French : synchronic and diachronic aspect*. Unpublished dissertation. University of Bradford.

Grant, A. P., 1995b, 'Article agglutination in Creole French: a wider perspective'. In P. Baker et C. L. R. Group (eds.), *From contact to Creole and beyond*, Westminster Creolistics series. University of Westminster Press.
Hagemeijer, T., 2009, 'Initial vowel agglutination in the Gulf of Guinea creoles'. In Enoch Aboh et Norval Smith (eds.), *Complex processes in new languages*. Amsterdam/Philadelphia: John Benjamins.
Ladhams, J., 2012, 'Article Agglutination and the African contribution to the Portuguese-based Creoles'. In Angela Bartens et Philip Baker (eds.). *Black through White*. London: Battlebridge Publications.
McWhorter, J., 1998, 'Identifying the creole prototype: Vindicating a typological class'*, Language* 74 -4, p 788–818
McWhorter, J. et Parkvall, M., 2002, 'Pas tout à fait du français, une étude créole'. *Études Créoles*, XXV-1, p. 179–231.
Miller, P., 1992, *Clitics and constituents in Phrase Structure Grammar*. New York: Garland.
Morin, Y-C., 1982, 'De quelques [l] non étymologiques dans le français du Québec: notes sur les clitiques et la liaison.' Revue québécoise de linguistique, 11-2, p. 9-47
Natchoo, Nicolas (Coordinateur) *Ki pase la? Bann Lavantir Vanessa ek Leo*. Std 2 (Volim 1). MIE Publications.
New B., Brysbaert M., Veronis J., et Pallier C., 2007. 'The use of film subtitles to estimate word frequencies'. *Applied Psycholinguistics* 28, p. 661–677.
New, B. et Spinelli, E., 2013. 'Diphones-fr: A French database of diphones positional frequency'. *Behavior Research Methods, 45(3),* p. 758-764
Plénat, M., 2000, 'Quelques thèmes de recherche actuels en morphophonologie française'. *Cahiers de lexicologie*, 77, p. 27–62.
Parkvall, M., 2008, 'The simplicity of creoles in a cross-linguistic perspective'. In *Language Complexity: Typology, contact, change*. Matti Miestamo, Kaius Sinnemäki et Fred Karlsson (eds). Studies in Language Companion Series, 94, p. 265–285.
Plénat, M., 2000, 'Quelques thèmes de recherche actuels en morphophonologie française'. *Cahiers de lexicologie*, 77, p. 27–62.
Robinson, S., 2008, 'Why pidgin and creole linguistics need the statistician'. *Journal of Pidgin and Creole Languages,* 23, p. 141–146.
Strandquist, R. E., 2003, *Article Incorporation in Mauritian Creole*, Master Thesis. Ph.D. thesis, University of Victoria.
Valdman, A., 1978, 'On the structure and origin of Indian Ocean Creole'. *Romance Philology* 32, p. 65-93.
Veenstra, T. (ms). 'Grammatical Aspects of the Bantu Diaspora'. RMCA, Tervuren, Feb 2011.
Wittman, H. et Fournier R., 1982, 'L'agglutination nominale en français colonial.' *Revue québécoise de linguistique théorique et appliquée 2 : 2*, p. 185-209.
Zribi-Hertz A., et Loïc J-L., 2013, 'From Noun to Name: definiteness marking in Modern Martinikè'. In Patricia Cabredo Hofherr and Anne Zribi-Hertz (eds). *Crosslinguistic studies on Noun Phrase structure and reference*. Brill, Syntax and Semantics, 39 : 269-315
Zwicky, A. N., 1977*, On clitics*. Indiana University Linguistics Club Publication.