

Éléments pour une analyse outillée des langues créoles

Emmanuel Schang

▶ **To cite this version:**

Emmanuel Schang. Éléments pour une analyse outillée des langues créoles. Linguistique. Université Paris Diderot, 2019. tel-03601171

HAL Id: tel-03601171
https://hal.science/tel-03601171

Submitted on 8 Mar 2022

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.

Éléments pour une analyse outillée des langues créoles

Emmanuel Schang

Document pour l'Habilitation à Diriger des Recherches
(Volume 1 : Document de synthèse)

Ecole doctorale 622 Sciences du Langage
Laboratoire de linguistique formelle, UFR de linguistique
UNIVERSITÉ PARIS DIDEROT
Le 9 décembre 2019

Jury composé de :

— Anne ABEILLÉ, Professeure des Universités (Univ. Paris Diderot), garante
— Olivier BONAMI, Professeur des Universités (Univ. Paris Diderot), président
— Viviane DÉPREZ, Directrice de Recherches CNRS (ISC & Rutgers University), examinatrice
— Claire GARDENT, Directrice de Recherches CNRS (LORIA), examinatrice
— Jean-Louis ROUGÉ, Professeur des Universités (Univ. Orléans)

Remerciements

Mes remerciements vont naturellement vers mes collègues du département des Sciences du Langage d'Orléans et du Laboratoire Ligérien de Linguistique, et parmi eux, je pense particulièrement à Jean-Louis Rougé avec qui j'ai eu beaucoup de plaisir à réaliser des travaux sur les créoles portugais d'Afrique. Je souhaite aussi remercier tous ceux qui ont été des partenaires importants durant ces années. J'en oublie certainement, qu'ils me pardonnent, mais je pense surtout à Yannick Parmentier, Simon Petitjean, Denys Duchier, Anaïs Lefeuvre-Halftermeyer à l'Université d'Orléans ; Jean-Yves Antoine, Denis Maurel et Agata Savary à Tours et Blois ; à Anne Zribi-Hertz, Patricia Cabredo-Hoffherr, Alain Kihm et plus généralement, les membres du GDRI SEEPiCLa et du Groupe de Recherche sur les Grammaires Créoles de SFL.

Enfin, je tiens à remercier particulièrement :

— Anne Abeillé pour avoir accepté d'être marraine de cette HDR,
— les rapportrices Claire Gardent et Viviane Déprez,
— ainsi qu'Olivier Bonami et Jean-louis Rougé qui ont accepté d'être membres du jury.

Table des matières

1 Introduction **7**

1.1 La linguistique de terrain 8

1.1.1 L'avikam 9

1.1.2 São Tomé et Príncipe 10

1.1.3 La Guyane 13

1.1.4 La Guadeloupe 13

1.2 L'outil informatique pour la linguistique de terrain 13

2 Evolution **15**

2.1 Pidgin, Créole et Décréolisation 16

2.1.1 Définir les notions 17

2.2 Créolistique 24

2.2.1 Phonologie des créoles du Golfe de Guinée 25

2.2.2 La profondeur historique 27

2.2.3 Les Angolares, le naufrage et les gènes 29

2.2.4 Conclusion partielle 33

2.3 GDRI SEEPiCLa 34

3 Structure **37**

3.1 Quel est l'objet d'étude ? 38

3.1.1 Continuum ou diglossie ? 38

3.1.2 Corpus de données de terrain 39

3.2 Grammaires électroniques 40
3.3 TAG pour la créolistique : motivations 45
3.4 Contribution 47
3.4.1 Utiliser une métagrammaire 47
3.4.2 TMA 50
3.4.3 Les prédicats non verbaux 61
3.4.4 La négation 66
3.4.5 Les articles 73
3.4.6 Les séries verbales 80
3.4.7 Conclusion 81

4 Analyses sur corpus **83**
4.1 Anaphores et coréférence 83
4.1.1 Retour en arrière 83
4.1.2 ANCOR 85
4.1.3 ANCOR-971 88
4.1.4 Contribution à l'analyse des chaînes de référence en créole 94
4.2 Temporalité 102
4.3 Intonation, prosodie 105
4.3.1 Tons, intonation et créoles 105
4.3.2 Question under discussion 108

5 Perspectives **113**
5.1 Constitution de corpus de large taille sur les langues créoles . . . 113
5.2 Apprentissage automatique au service de la linguistique 114
5.3 Métagrammaires 116
5.4 Le mot de la fin 118

Chapitre 1

Introduction

Sommaire

1.1 La linguistique de terrain 8

1.1.1 L'avikam 9

1.1.2 São Tomé et Príncipe 10

1.1.3 La Guyane 13

1.1.4 La Guadeloupe 13

1.2 L'outil informatique pour la linguistique de terrain 13

Dans ce mémoire pour l'obtention de l'Habilitation à Diriger les Recherches, je vais présenter les principaux travaux que j'ai réalisés depuis l'obtention de ma thèse. Je ne détaillerai pas le contenu des articles et des différents travaux mais je présenterai le contexte dans lesquel ils s'inscrivent. Je tenterai de donner une cohérence à cet inventaire. Nous le verrons, mes travaux vont dans diverses directions. L'explication à cela tient essentiellement dans le mode de fonctionnement de l'enseignement et de la recherche en France : la recherche de financements occupe une place prépondérante dans l'activité d'un enseignant-chercheur. Toutefois, le cap que je suis, même si ma navigation prend la forme de cabotage, retse le même et ma motivation reste intacte : c'est la description des langues peu connues et peu décrites qui m'a attiré et continue de m'attirer vers la linguistique.

Je continue de me considérer comme un linguiste de terrain, même si l'utilisation d'outils informatiques me pousse vers le traitement automatique des langues naturelles (TAL). Voyons le côté positif cependant : le montage de projets, la réflexion qui s'y rapporte, la constitution d'équipes pour relever des défis et le partage de connaissances qui y sont associés, tout cela est aussi une véritable richesse. J'ai eu la chance de coordonner plusieurs projets : un appel à projet régional (APR-IA) sur les anaphores (ANCOR) ainsi que le réseau de recherche international SEEPiCLa notamment. Cela m'a permis de mieux comprendre l'organisation de la recherche en France, mais aussi dans d'autres pays (le GDRI SEEPiCLa rassemble des unités des USA, Allemagne, Portugal, Royaume-Uni, Pays-Bas, Haïti, Maurice au côté d'UMR françaises). Et surtout, c'est une expérience humaine très riche.

Ce mémoire s'organise en trois parties qui se répondent :

— la première partie, intitulée "évolution" porte sur mes travaux en lien avec l'évolution et l'émergence des langues créoles.
— la seconde, "structure", porte sur la syntaxe essentiellement et sur la description des langues créoles et des langues peu dotées.
— la troisième est consacrée à l'"analyse de corpus". La mise à disposition de corpus est un enjeu majeur pour les langues 'peu dotées'. Recueillir des données authentiques et les rendre disponibles, mais aussi développer des outils d'analyse est un défi passionnant. Cela me semble justifier une partie entière dans ce mémoire.

1.1 La linguistique de terrain

Ce paragraphe a pour seule prétention d'éclairer le lecteur sur mon itinéraire en essayant de lui faire comprendre "d'où je parle".

1.1.1 L'avikam

Mes études de linguistique ont débuté après mon service militaire, lorsque j'ai décidé de reprendre des études universitaires. J'ai entrepris une maîtrise de linguistique à l'université Nancy 2 sous la direction du Professeur H-C Grégoire, africaniste et ancien directeur de l'Institut de Linguistique Appliquée d'Abidjan (ILA) entre autres fonctions. J'ai souhaité étudier des langues peu décrites et H-C Grégoire m'a indiqué les langues lagunaires de Côte d'Ivoire. En effet, celles-ci n'avaient fait l'objet que de descriptions succintes : quelques articles, quelques pages dans l'Atlas des Langues Kwa de Côte d'Ivoire et quelques mémoires de maîtrise à l'Institut de Linguistique Appliquée d'Abidjan. Mon choix a porté sur l'avikam. Cette découverte de la linguistique de terrain et des méthodes d'enquête s'est faite lors d'un séjour de quelques semaines à Abidjan, au centre de la SIL, et sur le terrain dans le village de Braffédon (Grand-Lahou) en compagnie de Jacques Rongier. J'ai pu bénéficier des séminaires hebdomadaires de la SIL et des séminaires de recherche à l'ILA. L'accueil extrêmement bienveillant que j'ai reçu à Abidjan et à Braffédon m'ont décidé à poursuivre dans cette voie : la linguistique descriptive. A mon retour, j'ai présenté un mémoire de maîtrise qui a été bien reçu par les enseignants de linguistique (19/20). Ce mémoire constitue d'ailleurs la référence pour l'avikam dans le Word Atlas of Linguistic Structures (WALS). J'aurais souhaité continuer la description des langues Kwa lagunaires mais les circonstances ont fait que je n'ai pas eu de bourse pour effectuer un DEA en Côte d'Ivoire. Par ailleurs, la situation politique de la Côte d'Ivoire commençait à se tendre et les chances de pouvoir effectuer un travail de longue haleine (une thèse) dans ce pays s'amenuisaient.

Ma compagne ayant eu l'opportunité de partir enseigner le français à São Tomé et Príncipe, j'ai décidé de la rejoindre pour effectuer une recherche sur les langues de cette île, assez peu connues à cette époque. On peut rappeler qu'en 1995, l'internet n'en était qu'à ses débuts et que les documents se commandaient en prêt inter-universitaire, ce qui rendait la rédaction d'un état de l'art longue et fastidieuse, en comparaison avec l'époque actuelle.

1.1.2 São Tomé et Príncipe

Avant de partir pour São Tomé, j'ai chercher à contacter Jean-Louis Rougé qui, à ma connaissance, était le seul linguiste français à avoir travaillé sur cette île. A mon retour de mission, Jean-Louis Rougé m'avait gentiment envoyé une lettre m'indiquant qu'il était au Cap-Vert et qu'il s'occupait désormais essentiellement de la formation des enseignants capverdiens. Ce n'est que lorsque Jean-Louis Rougé a été affecté à un poste en France que nous avons commencé à travailler ensemble. Contrairement à ce que beaucoup pensent, je n'ai pas été encadré pour ma thèse par Alain Kihm, qui aura cependant la gentillesse d'être un des rapporteurs de ma thèse, mais par Henri-Claude Grégoire qui, bien qu'africaniste, a consenti à encadrer mon travail. Comme il le rappelait à l'occasion, São Tomé se trouve bien en Afrique et rien n'empêche les africanistes de s'intéresser aux créoles. Ceci aura une influence importante sur mon travail car je n'ai jamais envisagé les créoles comme des 'questions de créolistique' mais comme des langues que l'on doit décrire, observer et que l'on peut, dans un second temps, relier à d'autres langues dans une perspective diachronique d'évolution.

Cependant, il convient, dans un souci de vérité scientifique autant que par auto-dérision, de revenir sur le motif initial de mes travaux sur São Tomé : décrire cette langue bantoue appelée 'ngola'.

Avant de partir, j'avais pu obtenir une page A4 photocopiée contenant une liste de phrases (proverbes principalement) de la lingwa ngola, que j'identifiais comme étant cette fameuse langue bantoue. J'avais repéré dans ces phrases des numéraux bantous (kwana 'quatre', samano 'cinq', etc.) et des mots d'origine bantoue (mijonga 'mer', tetembu étoile'), ce qui m'avait conforté dans cette idée de sujet de DEA. Cependant, arrivé sur place, il m'a fallu constater que ce ngola n'était pas une langue bantoue, mais un créole portugais comprenant une forte proportion de mots d'origine bantoue dans son lexique courant. Les livres sur la question, Maurer (1995) et Lorenzino (1998), ne sortiraient que plus tard. Cette anecdote a certainement contribué à forger mon caractère sur deux points :

— une force de conviction : lorsqu'il a fallu annoncer à mon directeur (de

DEA à l'époque) que j'avais passé trois mois à chercher une langue bantoue qui n'existe pas et que mon mémoire allait porter sur une langue créole,

— dans ma conviction saint-thomasienne (comme São Tomé, est-ce un hasard ?) de ne croire que ce que je vois.

Mon entrée dans la créolistique s'est donc faite sur un quiproquo...

Plus sérieusement, mes travaux à São Tomé ont été fortement influencés par les membres des organisations qui ont largement aidé à la réalisation des enquêtes de terrain, parmi lesquelles :

— les Volontaires du Progrès (VP),
— la Caisse Française de Développement (CFD),
— l'ONG Alisei, ex Nuova Frontiera (NF),
— le Centre d'Enseignement du Français.

Bien que je n'aie jamais été membre de ces organisations, j'ai reçu leur appui plus ou moins indirectement, soit par les facilités de logement accordées, par les facilités de transport, les contacts avec les autorités et les communautés.

Je pense notamment à l'énorme appui apporté par Tiziano Pisoni (NF) pour les contacts avec les petites communautés de Boa Morte, de Caixão Grande, de l'île de Príncipe mais aussi à la CFD qui m'a associée à ses travaux sur le microcrédit dans les communautés angolares. Leur aide m'a été précieuse pendant ma thèse, mais aussi après, et même lors de la mission financée par le CORAL (qui est devenu le Laboratoire Ligérien de Linguistique par la suite) en 2005.

J'ai trouvé au contact de ces personnes engagées auprès des populations locales un intérêt certain pour les langues créoles. Les besoins exprimés par ces ONG et par les populations m'ont incité à sortir du cadre de mes questionnaires grammaticaux pour aller vers la pragmatique, l'analyse du discours (très modestement) et une sociolinguistique ancrée dans la réalité (quel est le "marché des langues" ?).

Concrètement, cela s'est traduit principalement par des travaux de 'localisation' de documents (utilisation de termes créoles dans des documents). Par

exemple, j'ai participé à l'élaboration de l'expression *kai lokê lelu* ('maison fermer argent') pour désigner le concept de 'caisse d'épargne'.

Sans cet appui des ONG, je n'aurais jamais pu effectuer ce travail de terrain. Plus le temps passe, plus il me semble indispensable de connaître le terrain sur lequel les langues que l'on étudie sont parlées. Mon collègue Jean-Louis Rougé me rappelle souvent la parole de Mao Zedong (dans 'Contre le culte des livres' en 1930) : "Celui qui n'a pas fait d'enquête n'a pas droit à la parole". Je prends ici ces mots au sens littéral, au sens de l'enquête de "terrain".

Du point de vue de la créolistique, São Tomé et Príncipe revêtent une importance particulière. En effet, les créoles portugais du Golfe de Guinée font partie des premières langues reconnues comme 'créoles' et leur éloignement avec le Portugal les rend très intéressantes. En effet, à côté du créole santomense (ou forro, improprement appelé par certains 'santome') et qui est parlé dans la majeure partie de l'île, on trouve un autre créole, dont je viens de parler, qui s'appelle l'angolar (ou lungwa ngola) parlé à l'origine par les esclaves marrons. A côté des langues créoles existent les langues des Tongas (dont la première mention est Rougé (1992)). Il s'agit du 'portugais des Tongas', une variété de portugais en contact avec des langues bantoues, et des vestiges de langues bantoues restructurées, issues du Mozambique ou d'Angola. J'ai eu l'opportunité de travailler sur une variété de kimbundu avec l'un des derniers locuteurs de cette variété. Quant à Príncipe, on peut dire que le créole de l'île, le lung'Ie, est quasiment éteint. Des chercheurs brésiliens travaillent avec les derniers locuteurs de ce créole (v. dos Santos Agostinho (2016)). On peut déplorer la standardisation imposée par les moyens modernes de communication, mais il faut souligner que grâce à Internet, de nombreux santoméens émigrés au Portugal ou, plus rares, en France, peuvent désormais écouter des chansons en angolar. Espérons que cela permettra de sauver la culture et la langue angolares.

1.1.3 La Guyane

J'ai eu la chance d'effectuer une mission de deux semaines (financement LLL) à Saint-Georges-de-l'Oyapock en 2003 en compagnie de mon collègue Jean-Louis Rougé. Nous avons travaillé sur les langues en contact dans cette commune située à la frontière avec le Brésil. Le point le plus intéressant pour nous aura été le travail sur un exemple de variété régionale du français en contact avec d'autres langues locales : le portugais, le créole français de Guyane et le palikur (langue arawak). Nous avons ramené des enregistrements que nous exploitons dans nos cours respectifs et dont nous faisons mention dans nos travaux (v. Schang (2004) par exemple).

1.1.4 La Guadeloupe

En raison des difficultés objectives à concilier le travail de terrain et le travail d'enseignant-chercheur mais aussi à obtenir un financement de mes recherches de terrain, j'ai peu à peu quitté le terrain des créoles portugais pour commencer à étudier le créole de la Guadeloupe, île sur laquelle j'ai souvent l'occasion de me rendre à titre personnel (et sur des financements personnels). Cette reconversion a pu se faire aisément en raison de ma connaissance du terrain, de nombreux locuteurs et de mon expérience sur d'autres terrains. Je n'aborde cependant pas la question de la genèse de ce créole car je ne pense pas pouvoir apporter plus que les contributions antérieures Hazaël-Massieux (1996); Prudent (1999); Chaudenson (2004); Hazaël-Massieux (2008) notamment.

1.2 L'outil informatique pour la linguistique de terrain

Dans la mesure où l'on aura bien compris que c'est la linguistique de terrain qui m'a attiré avant tout, on peut s'étonner de voir dans mes travaux des articles

qui ont plus de lien avec le traitement automatique qu'avec la description 'classique' des langues.

Cela s'explique, dans un premier temps, par la difficulté d'obtenir des financements pour les travaux de terrain. Nécessité faisant loi, j'ai eu l'opportunité de financer mes études par des vacations au LORIA et à l'ATILF à Nancy. Les travaux sur corpus m'ont permis de me familiariser avec l'outil informatique et la linguistique de corpus.

Mais, dans un second temps, j'ai trouvé dans l'informatique et la linguistique de corpus des techniques et une méthodologie intéressante que j'ai souhaité (voire rêvé) appliquer aux langues sur lesquelles j'ai travaillé.

Chapitre 2

Evolution

Sommaire

2.1 Pidgin, Créole et Décréolisation 16

2.1.1 Définir les notions 17

2.2 Créolistique 24

2.2.1 Phonologie des créoles du Golfe de Guinée 25

2.2.2 La profondeur historique 27

2.2.3 Les Angolares, le naufrage et les gènes 29

2.2.4 Conclusion partielle 33

2.3 GDRI SEEPiCLa 34

Dans les lignes qui suivent, je présente rapidement (et donc forcément de manière partiale) le champ de la créolistique afin de pouvoir discuter les points qui me paraissent poser problème dans la section 2.2 et de situer mes travaux dans un ensemble plus vaste de travaux qui se répondent. Dans la mesure où de nombreux livres et articles font l'inventaire des théories sur la genèse des langues créoles, je m'épargne la répétition de cette tâche et je renvoie le lecteur aux ouvrages suivants : Velupillai (2015); Holm (1989); Mufwene (2005); Bakker et al. (2017); Romaine (2017); Mühlhäusler (1986); Valdman (1977) notamment.

2.1 Pidgin, Créole et Décréolisation

Si les études créoles trouvent probablement [1] leur origine dans les travaux de Schuchardt (Kreolische Studien) et de Hesseling à la fin du 19e siècle, il faudra attendre les années 60 pour qu'émerge un champ d'investigation nouveau que je vais appeler ici "créolistique" même si le terme n'est pas encore consacré à cette époque. Comme en témoigne DeCamp dans l'introduction des actes de la conférence de Mona en 1968 (DeCamp (1971)), très peu de linguistes étudiaient jusque-là à la fois des pidgins et des créoles. Et encore moins nombreux étaient ceux qui allaient regarder hors d'une seule aire géographique (les Caraïbes par exemple) ou en dehors d'une famille linguistique (les langues romanes par exemple). Ainsi, le pidgin anglais de Chine et le créole jamaïcain n'étaient jamais rassemblés au sein d'un seul champ d'investigation. C'est au tout début des années 60 que l'on commence vraiment à parler de "pidgin-creole studies" et de créolistes, même si Schuchardt avait déjà brièvement exposé l'idée d'une évolution des pidgins vers les créoles dans Schuchardt (1914). Et c'est certainement les théories monogénétiques qui ont favorisé l'essor de ce nouveau champ. Les définitions proposées par DeCamp (1971), même si elles sont contestées par les spécialistes, ont fait leur chemin dans la linguistique générale et se sont imposées. Il me semble qu'aucun linguiste aujourd'hui ne nierait l'importance de la créolistique au sein de la linguistique, quitte parfois à lui accorder une place un peu exagérée. C'est notamment, il me semble, le cas de Pinker (1999) qui à la suite de Bickerton (1981) voit dans les créoles l'occasion de découvrir comment des individus créent une langue à partir de rien :

> "Or le linguiste Derek Bickerton a démontré que, dans bon nombre de cas, un pidgin peut se transformer d'un seul coup en un langage complexe complet : il suffit qu'un groupe d'enfants soit exposé au pidgin

1. Dans l'introduction au fameux Valdman (1977), John Reinecke (1904-1982), qui a connu les créolistes du début du 20e siècle, défendait une histoire assez différente de celle de DeCamp. Il rappelait que les langues créoles étaient connues et faisaient l'objet de descriptions au moins depuis la *Grammatica over det Creolske Sprog* de J. M. Magen publié en 1770.

> à l'âge où ils acquièrent leur langue maternelle." (Pinker (1999)[p. 31])

Claude Hagège fera d'ailleurs, et à juste titre, la critique de cette vision du 'laboratoire créole' dans Hagège (1996).

Bien que je ne partage pas, on le verra, les analyses proposées dans les ouvrages cités ci-avant, il me semble que leur principale vertu était d'exposer aux yeux d'un public assez large les langues créoles, faisant sortir ces dernières des 'patois' et autre 'petit-nègre' [2]. D'une certaine manière, même si les créolistes ont parfois des débats un peu trop virulents (le débat scientifique a besoin de pondération, il me semble), ceux-ci sont probablement préférables à l'absence de prise en compte des créoles dans la linguistique générale.

2.1.1 Définir les notions

Pidgins

On trouvera de nombreuses définitions des langues pidgins et créoles et autant d'histoires différentes de la créolistique. Cette diversité de points de vue s'explique par l'absence de consensus sur la génèse des langues créoles. Mufwene (2001b) souligne le fait que les langues lexificatrices des créoles ont été à juste titre identifiées comme des **koinés** coloniales. On peut définir cette koiné comme un dialecte colonial construit avec l'apport de différents dialectes de la langue lexificatrice.

Mühlhäusler (1986)[chap1] passe en revue plusieurs définitions des **pidgins** :

— "une variété dont la grammaire et le vocabulaire sont très réduits… la langue résultante doit n'être la langue maternelle de personne" chez Bloomfield (1962).

— l'UNESCO définit un pidgin comme une langue née du contact entre personnes de langues différentes et habituellement formé par le mélange (mixing)

2. Meijer and Muysken (1977) consiste en une mise en perspective des travaux de Schuchardt et Hesseling dans le contexte de l'époque, soulignant très bien les préjugés racistes (et colonialistes) que l'on pouvait trouver partout alors.

des langues.

— une forme dont la structure a été très simplifiée bien plus que ce que l'on peut trouver dans les langues impliquées dans la formation de ces pidgins, pour Jespersen (2013).

On le voit, les définitions (il y en a bien d'autres...) insistent sur le caractère 'simplifié' des pidgins. Ceci aura une importance cruciale dans les travaux sur la créolisation. Partant de ces notions de réduction et de simplification, les travaux ont été orientés dans une direction dont les créolistes peinent à dévier.

Velupillai (2015) passe en revue les différents types de pidgins. Du pidgin 'stable' au pidgin 'étendu' (aussi appelé *pidgincreole*), toute une variété de formes et de situations sociolinguistiques existent :

— les pidgins de commerce et de les pidgins nautiques,
— les pidgins 'd'employés' (workforce pidgins),
— les pidgins de plantation,
— les pidgins des mines et des industries,
— les pidgins militaires,
— les pidgins urbains.

Cet inventaire montre bien les différents contextes d'utilisation des pidgins, mais ne dit rien des structures linguistiques en question.

Dès lors que l'on imagine qu'il y a une catégorie homogène et distincte appelée 'pidgin', on se heurte aux différences de structure entre ces pidgins. Ainsi Velupillai (2015) met bien en évidence les differences entre les structures sociales au sein desquelles naissent les pidgins et pour lesquelles il est difficile de trouver un dénominateur commun.

La définition d'une langue comme pidgin (et ses différents degrés) ne va donc pas de soi et les conférences de la SPCL et de l'ACBLPE sont souvent le lieu de débats sur la catégorisation pidgin/extended-pidgin/pidgincreole/créole/semi-créole. On peut voir cela comme le signe que cette question n'est pas résolue ou bien qu'elle est mal posée.

Créoles

Mais quel est l'enjeu pour la créolistique derrière la catégorisation 'pidgin' ou 'créole' ?

On trouve dans Bakker et al. (2017) une explication synthétique de la question de la créolisation :

> The traditional view of pidgin and creole genesis (as laid out for example in Mühlhäusler 1997 ; Romaine 1988) is that pidgins came about in contact situations where groups of people without a common language had to communicate with each other, and crude, simplified contact languages were created, first quite individual and normless jargons, later more systematic and collective pidgins (on pidgins, see Parkvall & Bakker 2013). When these simplified languages had to be used for other functions and when they were employed as languages of wider communication, or became mother tongues (nativization), the limited structures and lexicons of pidgins were expanded, thus developing into languages of full communicative functionality. Those fully developed languages are called creoles and the process of the development of jargons or pidgins into creoles is called creolization. Creoles are almost always native languages or mother tongues, in contrast to pidgins. Pidgin-derived contact languages that are spoken as both second and first languages are sometimes called *expanded pidgins* or *pidgincreoles* (Bakker 2008), and due to their functional and structural expansion they are quite similar to creoles (e.g., Tok Pisin, Nigerian Pidgin English, Sango, Plains Indian Sign Language). [p.6]

Mühlhäusler (1986) présente une analyse très complète [3] des différentes théories sur la créolisation et compare les différents arguments avancés à l'appui des différentes thèses. Il prend garde à ne pas caricaturer les positions et débute le

3. Voir également Holm (1989).

chapitre consacré aux origines des créoles (chap. 4) par une réserve :

> I shall start by discussing each of these theories in isolation, although it seems unlikely that any single cause will be sufficient to explain the complex processes which call these languages into being.

Les théories évoquées sont donc :

— l'approche basée sur les jargons nautiques (autour de Reinecke notamment, avec comme point de départ une remarque de Schuchardt),
— les théories du *baby talk* et du *foreigner talk*,
— la relexification,
— diverses théories substratistes,
— l'influence du supertstrat.

L'un des scénarios les plus souvent exposés (v. Romaine (2017) par exemple qui consacre le chapitre 5 de son ouvrage au *cycle de vie des créoles : décréolisation et recréolisation*) est celui de la naissance des créoles (1) par expansion :

(1) Type 1 : jargon > créole
Type 2 : jargon > pidgin stabilisé > créole
Type 3 : jargon > pidgin stabilisé > pidgin étendu > créole

D'une certaine manière, la théorie universaliste de Bickerton Bickerton (1981) appartient à la même veine.

Je ne vais pas reprendre les nombreux contre-arguments opposés à ces approches. Le lecteur pourra se référer à Chaudenson (2004) et Mufwene (2005) notamment.

Il apparaît cependant que malgré les critiques sévères qui peuvent être faites, ce modèle semble toujours connaître un certain succès (v. notamment Pinker (1999, 1995); Hinzen (2006) qui le reprennent sans aucune distance, comme je l'ai déjà montré précédemment). Bakker et al. (2017); McWhorter (2011) sont des exemples récents de travaux dans ce cadre. Bien que très contesté, les auteurs le reconnaissent, ce scénario connaît un succès que des études plus fouillées peinent à concurrencer. De ce fait, j'appelerai les scénarios de formation des créoles qui

font appel à cette idée d'expansion à partir d'une forme réduite, peu importe son origine, le *modèle standard*.

Sur quoi repose ce modèle ?

Le cycle pidgin > créole > décréolisation, idée dominante dans le champ de la créolistique, mérite que l'on s'y attarde un peu.

L'idée principale est celle du continuum créole, et la notion de *décréolisation* qui l'accompagne. Bickerton (1981) notamment, et suite à DeCamp (1971), décrit la décréolisation comme un processus qui apparaît lorsqu'un créole est au contact avec son superstrat.

On trouve d'un côté de ce continuum le *basilecte* (variété la plus éloignée de la langue lexificatrice) et à l'autre pôle, l'*acrolecte*. Entre les deux se situent les variétés *mesolectales*.

Le 'vrai' créole est la variété basilectale, celle qui se situe dans un écart maximal avec le superstrat.

Cette idée trouve un écho chez les défenseurs de la langue créole, comme en témoigne cet extrait d'un document de Jean Bernabé (http://www.touscreoles.fr/) :

> Le terme peu connu dans le grand public de « décréolisation » désigne un processus ayant rapport à la perte, au délitement, à la désagrégation du créole. Cette question, qui n'est pas toujours prise au sérieux à la mesure de ses enjeux, renvoie à une problématique que je considère comme cruciale pour l'avenir linguistique , culturel, voire politique (au sens noble du mot) de nos pays. En effet, le mot « décréolisation » constitue une métaphore particulièrement pertinente de la situation dans la quelle se trouvent nos pays, au carrefour d'un consumérisme débridé et d'une productivité voisine de zéro. La langue peut, en effet, elle aussi être un lieu tant de consommation que de production. Il est évident que dans ce domaine les créolophones se situent dans une

> situation qui accuse une disparité entre consommation linguistique et créativité. [...]
>
> Sous sa forme qualitative, la décréolisation est ce processus qui entame la substance, la qualité même du créole. Cette décréolisation est en fait une francisation quand la langue de contact est le français. Mais elle peut tout aussi bien être une anglicisation, quand, comme à la Dominique et à Sainte-Lucie, le contact s'établit avec l'anglais. Remarque importante : la décréolisation qualitative n'entraîne pas forcément la décréolisation quantitative, c'est-à-dire le recul ou la mort du créole. En effet, les Martiniquais parlent en général un créole très décréolisé, mais cela n'empêche pas leur créole d'être très vivace, comparé à certaines langues de France, comme l'occitan ou encore le breton, qui ne sont guère parlés que par des gens âgés. Il n'y a donc pas de raison que le créole martiniquais disparaisse parce qu'il se francise. Mais il n'y a pas non plus de raison de penser que ce créole ne disparaîtra jamais ! Divers facteurs sont à l'œuvre.

Entendu comme cela, la décréolisation est une question culturelle. Il est tout-à-fait légitime de défendre l'usage des formes créoles. Et l'on peut aisément comprendre que pour pouvoir enseigner ou promouvoir la langue créole, il est nécessaire d'établir des différences avec le superstrat, quitte à les accentuer. L'idée n'est donc pas à rejeter, elle a son utilité. Cependant, prise sans aucune distance critique, elle amène à des analyses fausses. J'y reviendrai dans la section suivante plus en détail en présentant mes travaux, mais je présente ci-après quelques arguments qui ne figurent pas explicitement dans mes travaux.

Qui parle le "basilecte" ?

Personne. Le basilecte est conçu dès l'origine comme une construction. Certains locuteurs employent des formes plus basilectales que d'autres à certains moments mais aucun locuteur ne parle le créole basilectal. On entend très souvent dans les conférences de créolistes des exposés qui manquent totalement de distance sur cette notion. Pourtant, le travail de terrain nous rappelle sans cesse les

faits : il n'y a pas de locuteur du basilecte.

Il y aurait beaucoup à dire de ce continuum créole, mais je me contenterai ici de noter que même si on accepte cette idée, rien ne nous oblige à suivre ce qui est proposé dans certains travaux.

En effet, identifier le créole comme rassemblant ce qui est le plus éloigné du superstrat renforce bien entendu l'idée que les créoles sont (nécessairement) issus de pidgins (entendus comme simplification). Le choix des traits linguistiques donnés/choisis comme présents dans les créoles est d'ailleurs un enjeu majeur :

— les bases de données de type APICS Michaelis et al. (2013), par le choix des traits les plus basilectaux ont un parti pris typologique fort,

— les études telles que Bakker et al. (2017); Bickerton (1981); McWhorter (2011) qui cherchent à mettre en évidence l'existence d'un type créole excluent de fait toute la complexité qu'intègre le continuum. Ceci a été déjà contesté par G. Fon Singh et J. Léoué à plusieurs reprises (Sing (2017)).

Mufwene (1997, 2001a) apporte des éléments convaincants au débat :

> "Quand les linguistes et plus particulièrement les créolistes, parlent de processus de *créolisation*, ils entendent par là le processus de *basilectalisation* (Chaudenson 1992, 2001, 2003, Mufwene 1996, 2001), c'est-à-dire, l'émergence d'un basilecte à partir de la langue coloniale européenne que les populations serviles se sont appropriées. Le *basilecte* est la variété qui, dans le continuum associé aux communautés créolophones, est structurellement la plus différente de la variété "standard" parlée par l'élite locale, appelée *acrolecte*. De même, les créolistes identifient comme *décréolisation* un processus qui devrait plutôt être appelé *débasilectalisation* (Mufwene 1994, 2001), et qui consiste en la perte des straits structurels associés au basilecte. Ils nous font alors croire que le vrai créole serait seulement le basilecte et que le *mésolecte*, le reste du continuum entre les deux pôles qui est parlé par la majorité de la population, serait le résultat de ce processus de "décréolisation". "

Ceci résonne avec cet extrait de Labov (1971) cité notamment dans Mühlhäusler (1986) :

> "The real creole is always in the process of receding over the horizon... and creole grammars tend to be normalized descriptions of an earlier phase of the language that no one is quite sure was ever spoken by anyone. "

Par ailleurs, lorsque l'on évoque la décréolisation, on prend souvent comme exemple le fait que les jeunes disent désormais X alors que les vieux disaient Y. Si le changement lexical est un processus que l'on constate dans toutes les langues vivantes, il est pris chez les créolistes comme un signe de décréolisation. Pourtant, comme le remarquait Hazaël-Massieux (1996) face aux allégations de décréolisation du lexique en guadeloupéen :

> "En ce qui concerne le créole de Guadeloupe, il suffit de montrer qu'il a à peu près la même liste [lexicale] que le créole de la Dominique (l'intercompréhension entre les deux îles est totale), or cette île a cessé d'être française depuis 1763, c'est-à-dire une époque où le créole était sans doute formé avec ue physionomie proche de celle d'aujourd'hui, mais à partir de laquelle on peut penser qu'en l'absence d'une société française dominante, il est improbable qu'il y ait eu la moindre décréolisation en direcion du français. "

On le voit donc, la question est loin d'être tranchée.

2.2 Créolistique

Un regard rapide sur mon travail depuis 1995, date de mon premier terrain à São Tomé, me fait réaliser que la question de la créolisation (envisagée comme processus général) n'a pas été l'objet de mon intérêt. En effet, à aucun moment je n'ai souhaité proposer une quelconque piste sur le processus de créolisation en général, mes travaux se limitant à donner quelques pistes sur l'émergence de

langues créoles *en particulier* : les créoles portugais du Golfe de Guinée. Dans ma thèse, j'essaye d'évaluer l'influence des différents facteurs en cause dans la genèse de ces langues, mais ces questions seront abordées plus en détail dans les travaux que j'ai effectués avec Jean-Louis Rougé.

2.2.1 Phonologie des créoles du Golfe de Guinée

Dans Schang (2003) j'aborde les contraintes phonotactiques à l'oeuvre dans la genèse des créoles du Golfe de Guinée. Cet article est le prélude à un autre travail, plus précis et plus abouti, en collaboration avec JL Rougé. Dans Rougé and Schang (2006), nous mettons en évidence le fait que la liquide / l / en ST (et plus largement dans les différents CGG) n'est pas le résultat d'un processus simple de conservation d'un segment de l'étymon portugais, mais que des processus complexes et diachroniquement distincts sont à l'oeuvre. En particulier, nous montrons que certains mots (y compris issus d'étymons non portugais) ont fait l'objet de modifications à l'initiale (apparition d'attaques complexes *kl, xtl*) à une date plutôt récente (19e siècle vraisemblablement). L'hypothèse "uniformaliste" (un seul processus à l'oeuvre pour tous les mots) n'est viable (v. Bhatt and Hagemeijer) qu'au prix de l'exclusion de nombreuses exceptions (navlega "naviguer" < P. navegar, klonvesa "discuter" < P. conversar). Ceci ne peut que reposer la question telle que l'abordons. Notons au passage qu'aucune contrainte exprimée dans la théorie de l'Optimalité (preservation, etc.) ne peut exprimer l'apparition des attaques complexes comprenant une liquide non étymologique et impliquant un marquage net. Toutefois, l'hypothèse que nous avançons (compatible avec celle de Ferraz (1987)) a parfois été interprétée de manière surprenante. En effet, on trouve dans cet exposé de Hagemeijer et Araujo, l'extrait suivant :

> "In his study of Santome, Ferraz (1987) describes the contexts in which liquids in the Portuguese etymon were deleted, maintained or metathesized (cf. examples in Tables 1, 2 and 3) and briefly compares these patterns to those found in the three sister creoles. Ferraz hypothesizes that the declustering patterns were a characteristic feature

> of the proto-language and that Santome developed the consonant+/l/ sequences at a later stage, subsequently to speciation of the proto-language. In a follow-up study, Rougé & Schang (2006) suggest that the liquid consonants were introduced due to the growing Portuguese presence from the mid-19 th century on, during the coffee and cacao boom. It is crucial to note, however, that documents from that period written in Santome show that the consonant+/l/ sequences were already well established at that time (e.g. Negreiros 1895 ; Schuchardt 1882)." source : http://alfclul.clul.ul.pt/clulsite/FACS_III-Booklet.pdf

Il est bien entendu que le milieu du 19e siècle se situe avant 1880 et qu'en 30 ans, avec un renouvellement important de la population, des changements de ce type ont largement eu le temps de se produire. Ce n'est pas la peine de développer plus loin ces arguments mais, au delà de l'anecdote, on peut remarquer toute la difficulté de prendre en compte la profondeur diachronique chez certains créolistes. Ce même constat a été fait dans Chaudenson (2004). Les créoles seraient donc les seules langues sans histoire (histoire interne, j'entends) ? La seule évolution possible est-elle nécessairement la décréolisation ? Non, bien entendu. Je n'invente rien ici cependant, la voie étant déjà tracée pour les créoles du Golfe de Guinée par Jean-Louis Rougé depuis au moins Rougé (2004). Mais celle-ci commence seulement à trouver un écho parmi les (trop rares) spécialistes des créoles portugais.

L'idée que les créoles actuels illustrent nécessairement et directement un parler commun antérieur (proto-créole) reste vivace dans les études sur les créoles portugais Cosme (2014); dos Santos Agostinho (2016); Bandeira (2017).

La situation est un peu différente pour ce qui concerne les créoles français, les travaux de Chaudenson et de Mufwene, entre autres, étant très largement compatibles avec nos analyses.

2.2.2 La profondeur historique

De manière générale, et contrairement à ce que ma thèse pourrait laisser penser par l'utilisation du modèle des Contraintes et Réparations qui applatit la dimension diachronique, les processus à l'oeuvre dans la créolisation ne sont ni homogènes ni fixés. En d'autres termes, le forro tel qu'il est parlé à l'heure actuelle n'est pas le forro parlé lors des premières années du 16e siècle (années de sa formation). Les témoignages des navigateurs sur l'existence d'une langue qui n'est pas le portugais ne nous permettent pas de dire qu'il s'agit du forro actuel. Comme nous l'avons montré dans Rougé and Schang (2006, 2012), on peut trouver les traces d'une évolution diachronique du créole qui échappe au cliché pidgin>créole>décréolisation. Dans Rougé and Schang (2013), nous l'exprimons d'ailleurs clairement ainsi :

> "Et d'abord et avant tout que nous apprend-elle [la comparaison des créoles portugais d'Afrique] de l'émergence non pas de tous les créoles du monde, mais plus modestement de l'ensemble de ces langues ?"

Et plus loin :

> "Dès lors, la comparaison ne nous amène pas à reconstruire le portugais mais une ou des variété(s) d'apprenants. Ces variétés ont constitué ce qui va devenir la base du créole en devenir. Les recherches sur l'acquisition des L2 au cours des dernières décennies nous ont appris ce qu'il y a d'universel dans ces processus."

Dans cet article, nous procédons à la comparaison des quatre créoles de manière assez classique pour mettre en évidence ce qu'est ce (ou ces) proto-créole(s) dont la plupart des auteurs (Hagemeijer (2011); Bandeira (2017)) parlent sans toutefois donner de contenu à cette notion. Les exemples que nous donnons montrent qu'on peut arriver à dégager, zone par zone, des points communs qui ressemblent à ce que l'on trouve dans les variétés d'apprenants.

Il me semble donc que l'origine des créoles du Golfe de Guinée est donc à rechercher non dans un pidgin virtuel, mais dans les contacts de populations à

l'oeuvre sur ces îles. La reconstruction historique (cf. Rougé and Schang (2013)) nous amène à dégager des traits qui sont compatibles avec la coexistence de variétés d'apprenants stabilisées (donc des langues). Mais à aucun moment nous ne donnons au terme proto-créole [4] autre chose qu'un contenu théorique et nous admettons que de multiples variétés aient pu co-exister.

Dans cet article, nous insistons sur le fait que le proto-créole que l'on peut reconstruire avec les méthodes classiques de la linguistique diachronique n'est pas le créole actuel, de même que ce qu'on peut reconstruire du proto-roman n'est pas l'italien ni le français. A aucun moment, le fait d'aboutir à une forme (reconstruite) qui correspond assez bien à ce qu'on trouve dans des variétés de base ne nous fait dire que le créole (ou plus précisément la grammaire du créole santomense) est une variété de base. L'hypothèse est plutôt que la "négociation" autour de la parole s'est faite autour de variétés d'apprenants (par nature variées) et que, comme pour toutes les langues (le français également, s'il est nécesaire d'insister encore une fois), ceci n'est qu'un des ingrédients. Par cette approche, nous disons clairement que ce qu'on peut reconstruire ne nous mène ni au bini (edo), ni au portugais, même si ces langues ont évidemment joué un rôle important dans la formation des créoles du golfe de Guinée. Cette approche nous permet d'ailleurs d'intégrer des évolutions issues des langues bantoues (dans Rougé and Schang (2006)} notamment), langues que la main d'oeuvre contractuelle (les Tongas) parlait. Parmi toutes les variétés disponibles à chaque moment, il y a négociation au sein de ce marché linguistique (je reprends ici le terme de Bourdieu).

Ce que l'on peut reconstruire et ce qui était parlé dans une situation où un standard n'avait pas encore nécesairement émergé sont deux choses distinctes.

Il convient également de rappeler que mettre en évidence des éléments qui peuvent être reconstruits n'induit pas forcément que la thèse d'un développement pidgin > créole soit validée. La complexité du fonctionnement des négations en Santomense (v. Hagemeijer (2007)) s'oppose clairement à cela. Et enfin, si l'on

4. Nous parlons d'ailleurs de pré-créole (Eléments de comparaison de la lunga ngola et de la lungwa santomé : sur la piste d'un pré-créole) dans les communications qui ont précédé cet article, et dans Schang (2003).

peut reconstruire des formes différentes pour chaque zone (Haute Guinée et Golfe de Guinée), cela contredit clairement l'idée d'un seul processus de formation à l'oeuvre dans les créoles.

Ces hypothèses me paraissent compatibles avec des théories du type "écologiste" (cf. Mufwene (2005)) même si nous accordons probablement une attention plus importante à l'apprentissage des langues secondes [5].

2.2.3 Les Angolares, le naufrage et les gènes

Rougé and Schang (2012) est un article qui a été écrit en réaction à la lecture d'un article dans Courrier International ((Díario de Notícias repris dans Courrier international du 20 décembre 2007) qui mettait en avant les succès de la génétique dans la découverte de l'origine de la population angolare (sud de l'île de São Tomé).

> Cet article porte sur l'histoire de l'angolar, une langue créole à base portugaise parlée sur l'île de São Tomé. Dans un premier temps, nous détaillons les arguments linguistiques qui permettent d'affirmer que l'angolar est issu d'un proto créole commun à l'autre langue de São Tomé, nommée forro. Dans un second temps, nous montrons que les études récentes basées sur la génétique des populations apportent peu d'arguments décisifs, contrairement à ce qu'on peut lire dans la presse, lorsqu'il s'agit d'expliquer l'origine de ces langues via l'origine des populations. En conclusion, nous soulignons les dangers liés à une interprétation trop hâtive de ces données génétiques.

Cet article de Courrier International s'appuie sur un article scientifique écrit par une équipe pluridisciplinaire comprenant des biologistes et un linguiste Coelho et al. (2008) . Bien entendu nous ne discutons pas les arguments issus de la génétique car ce n'est pas notre domaine de compétence. En revanche, quelques passages nous ont laissé perplexes. On trouve notamment :

5. N'étant pas un spécialiste de l'acquisition, je me contenterai de suivre les arguments convaincants apportés par Jean-Louis Rougé.

> "And yet, human populations derived from the Atlantic slaving process are natural laboratories that provide a unique opportunity for integrative studies aiming at the identification of key evolutionary determinants of current patterns of human cultural and biological variation"

et plus loin :

> "São Tomé (...) may be considered an excellent model for assessing the microevolutionary impact and biocultural implications of the slave trade"

Les populations issues du commerce d'esclave seraient donc un laboratoire naturel.

Qu'est-ce que cela permettrait de valider ?

> 'The close parallel with the spatial distribution of genetic clusters confirms that the geographic patterning of genetic variation in São Tomé is mainly determined by the dichotomy between Angolares and non-Angolares. Additional confirmation is provided by the pattern of pairwise genetic distances calculated between sampling F st locations. The significance of the divergence between Angolares (cluster A) and non-Angolares (cluster B) is further attested by the finding that, when compared with 11 additional African-derived samples using pairwise genetic distances, the two clusters were never grouped together in any bootstrap replication, showing a level of differentiation (Fst = 0.03) that is clearly above the average (Fst = 0.01).'

> 'The fact that individuals from São João dos Angolares and Santa Catarina show the highest proportions of ancestry in cluster A, in spite of the relatively great distance between these villages, indicates that the genetic differentiation of cluster A is more closely related to language than to geographic distance.'

Nous montrons que l'étude en question ne confirme ni n'infirme rien de plus que ce que la linguistique et les connaissances historiques ne savent déjà. Et surtout, dans la tentative de réhabiliter l'hypothèse coloniale du naufrage (les Angolares seraient issus du naufrage d'un bateau d'esclave) n'est en rien rendue crédible.

Mais surtout, ce papier est l'occasion de s'intéresser aux liens entre l'origine d'une population et la/les langue(s) qu'elle parle. Que désigne-t-on par 'angolar' ? Existe-t-il une angolarité génétique ?

Comme nous le soulignons dans l'article :

> Deux des artistes angolars les plus connus, les plus actifs dans la promotion de la culture angolare, de cette tradition de pêcheurs, avouent sans problème avoir des ancêtres non-Angolar.

Nous apportons la conclusion suivante :

> De tout cela, on peut conclure que l'étude de Coelho & alii (2008) n'invalide l'hypothèse d'un pré- ou proto-créole commun au forro et à l'angolar qu'au prix des présupposés suivants :
> — il y a une corrélation stricte entre parenté génétique et langue/culture
> — l'origine des langues est à calquer sur l'origine de leurs locuteurs,
> Nous avons vu que le premier présupposé ne tient pas. Il en va de même du second. Rien n'exclut le fait qu'un groupe homogène de départ (le founder group de Coelho et al. (2008)) ne soit en contact avec des esclaves en fuite (comme l'attestent les données historiques) et que la langue qu'ils pratiquent soient ce proto-/pré- créole. Rien n'indique non plus que ce soit les femmes enlevées dans les plantations qui aient apporté le contact avec le forro (comme le prétend l'article de Courrier International). Car on retrouve là un vieux mythe colonial, celui du viol des femmes par les ennemis, qui, s'il repose sur un fond de vérité parfois, n'est là que pour conforter la peur de l'ennemi (ici, l'esclave marron).

Cela nous a amené à discuter également la "construction" des langues de São Tomé dans Rougé and Schang (2012).

> Identificar, classificar, denominar as maneiras de falar não é um assunto reservado aos linguistas. Cada dia, os locutores avaliam as producões linguisticas dos outros : «Fala bem », « têm sotaque » , « não fala » ou « fala português », etc. Esses julgamentos mais do que linguísticos, são julgamentos sociais que dizem respeito à constituição de comunidades humanas, à hierarquização das mesmas, à identificação das pessoas como membro das comunidades. Bourdieu (1982) já mostrou a violência simbólica dessas classificações já que os julgamentos se dirigem tanto às maneiras de falar como aos locutores que são assim identificados, classificados, denomina- dos. Nas situações de contactos de linguas essas avaliações tomam uma acuidade mais revelante. Trata-se muitas vezes de determinar onde começa ou onde acaba uma língua. Quando essas situações de contactos são no mesmo tempo situações coloniais ou post coloniais atrás das delimitações linguísticas surgem as delimitações sociais. As representações ligadas às mesmas transparecem nos nomes escolhidos para designar as variedades linguísticas (sobre os nomes das línguas Tabouret-Keller (1997)). Disso, São Tomé e suas línguas constitui um exemplo perfeito. [6]

6. “Identifier, nommer et classer les façons de parler n’est pas un sujet réservé aux linguistes. Quotidiennement, les locuteurs évaluent les façons de parler des autres : ’il a un accent’, ’il parle portugais’, ’il ne parle pas portugais’... Ces jugements sont plus de nature sociale que linguistique et liés à la façon de construire des communautés humaines, à leur hiérarchie et à l’appartenance de locuteurs à ces communautés. Bourdieu (1982) a déjà montré la violence symbolique de ces jugements et classements dirigés autant sur la manière de parler que sur les locuteurs. Dans les situations de contact de langue, ces jugements sont encore plus importants. Il s’agit souvent de savoir où commence et où termine une langue. Quand ces situations de contact sont aussi des situations coloniales ou post-coloniales, derrière les barrières linguistiques apparaissent des barrières soicales. Ces représentations transparaissent dans la façon de nommer les langues (voir Tabouret-Keller (1997)). L’île de Sao Tomé et les langues qui s’y parlent constituent le parfait exemple de cette situation.”[ma traduction]

L'un des points importants de cet article est l'insistance sur le lien entre E- et I-language dans la construction des objets linguistiques :

> Qualquer que seja a perspectiva escolhida é de sublinhar que as línguas **não existem em si** mas são objectos construidos e do nosso ponto de vista, na construção desse objecto, tanto a escolha dos traços estruturais em cada variedade como a relação entre I-Language e E-Language tomam a maior importância, o que significa que dificilmente se pode separar os aspectos linguísticos, i.e a estrutura interna, dos aspectos sociais. [7]

Nous faisons le lien entre la situation dans les plantations de São Tomé avec les recherches que nous avons effectuées en Guyane et qui soulèvent des questions similaires abordées dans Schang (2004). Réifier une (ou un ensemble) de productions comme une variété et la nommer ne va pas de soi et le linguiste 'de terrain' endosse une grande responsabilité. La charge symbolique de l'objet qu'il nomme et construit n'est jamais sans impact sur les populations.

2.2.4 Conclusion partielle

J'ai présenté ici les quelques travaux que j'ai effectués dans le domain de la diachronie. Ce n'est pas là que réside l'essentile de mon travail car, on l'aura compris, il me semble que

1. la question de la créolisation (telle qu'elle est abordée majoritairement) me semble mal posée,
2. pour faire mieux, il faut absolument pouvoir rendre compte de la richesse des données.

7. "Quelle que soit la perspective choisie, il faut souligner que les langues n'existent pas en soi mais sont des objets construits. De notre point de vue, les traits structuraux choisis pour chacune des variétés comme illustrant le lien entre I-language et E-language prennent une importance cruciale. Cela signifie qu'on ne peut que difficilement séparer les aspects linguistiques (la structure interne) et les aspects sociaux."[ma traduction]

C'est donc naturellement que l'essentiel de mes travaux sur les créoles sont tournés vers l'analyse synchronique. En particulier, je défends depuis ma thèse (Schang (2000)) l'idée qu'une approche basée sur le lexique est certainement la plus adéquate. Cette approche lexicaliste se retrouve par la suite dans les travaux sur la grammaire TAG que je présente ci-après.

2.3 GDRI SEEPiCLa

C'est dans ce contexte que se placent mes travaux dans le cadre du GDRI "créole". Le Groupe de Recherche International (International Research Network) *Structure, Emergence, Evolution of Pidgin and Creole Languages* est un projet de 'réseautage' (dans le jargon des projets de recherche) financé par le CNRS, sur les langues pidgins et créoles. Il est issu des travaux menés dans le cadre du goupe de recherche sur les grammaires créoles (GRGC) autour de D. Véronique et A. Zribi-Hertz initialement, puis d'A. Zribi-Hertz et P. Cabredo-Hoffherr plus récemment. A l'initiative d'Anne Zribi-Hertz et d'Alain Kihm a été monté la version initiale de ce GDRI (2011-2015), dont la coordination a été confiée à Olivier Bonami.

Je reprends ici les quelqus lignes qui définissent ce projet :

> Pidgin-creole studies are now mature : they share a corpus of concepts, questions and research agendas about which all researchers agree. Among European scholars there seems to be a consensus that (a) generalist theories of pidgin-creole formation are basically correct ; (b) pidgin-creole studies should not, as a consequence, be separated from all other studies, the intertwining or parallel pursuit of which constitute language science ; (c) it is a no less legitimate endeavour to study every individual pidgin or creole language for itself, as one does Russian or Hixkaryana ; (d) the endangered language angle is an important one in pidgin-creole studies.
>
> The present research programme aims to build on this consensus in order to reach a much higher level of coordination. The network groups

> specialists from : France, Germany, the Netherlands, Portugal, the United Kingdom, Haïti, the Republic of Mauritius and USA. They already have a long experience of working together on an individual basis or through more or less informal research groups. One such group, that has endured for several years now, is the Groupe de Recherche sur les Grammaires créoles (GRGC) which brings together Dutch and French scholars for yearly meetings. The level of synergy thus reached as well as the informal network that unites individual scholars across Europe fully justifies, even demands, we believe, establishing a more permanent structure with sufficient resources for continuous and institutionalized scientific coordination, in a way that has never existed in Europe so far. This is the goal of the present "International scientific coordination network". (source : http://www.pidgins-creoles.cnrs.fr/presentation)

J'ai participé à ce réseau initialement en tant que membre individuel (le LLL n'étant pas encore une UMR), puis j'ai rédigé le dossier de renouvellement. Cela a consisté principalement à :

— redéfinir la liste des tutelles participantes (des laboratoires de recherche) en l'orientant un peu plus vers des pays créolophones : Université de Maurice et Université d'Etat d'Haïti.
— orienter les financements vers la construction de projets bi- ou multilatéraux (France-Portugal, France-Haïti, etc.).

Ainsi, le GDRI a pu supporter des missions de formation à Port-au-Prince, des missions à l'Université de Maurice ou des missions au Portugal, et il a également contribué à l'organisation de conférences internationales, directement ou indirectement par la prise en charge de missions, comme par exemple :

— le colloque du CIEC à la Guadeloupe en 2016,
— le colloque Formal Approaches of Creole Studies (FACS-5) à Lexington en marge de l'école d'été de la Linguistic Society of America.
— FACS-6 à Orléans (selon toute vraisemblance) à l'automne 2019.

On peut affirmer que le point commun de ce réseau est de prendre au sérieux la description des langus créoles dans toute leur complexité. En particulier, je pense pouvoir affirmer que personne dans ce réseau ne considère satisfaisants les travaux typologiques du type APICS ou les approches phylogénétiques basées sur un inventaire de traits, comme dans Bakker et al. (2017). Les publications telles que Aboh (2015, 2016) et Sing (2017) [8] pointent clairement les problèmes méthodologiques des travaux de Bakker et al. (2017) et apportent de solides arguments contre l'idée d'un "type créole".

Les travaux du groupe montrent qu'il est nécessaire de passer rapidement à une phase de collecte de données de grande envergure sur les créoles (corpus oraux principalement) de façon à pouvoir disposer de matériaux riches pour l'étude des langues créoles. Le fait que des créolistes de premier plan travaillent, enseignent et sont des locuteurs créoles rend ce travail possible désormais. C'est, à mon avis, la tâche principale des créolistes pour les cinq années qui viennent.

8. Silvia Kouwenberg prépare également un article qu'elle a présenté au séminaire du GRGC qui remet fortement en question les analyses de Bakker et al. (2017).

Chapitre 3

Structure

Sommaire

3.1 Quel est l'objet d'étude ? 38
3.1.1 Continuum ou diglossie ? 38
3.1.2 Corpus de données de terrain 39
3.2 Grammaires électroniques 40
3.3 TAG pour la créolistique : motivations 45
3.4 Contribution 47
3.4.1 Utiliser une métagrammaire 47
3.4.2 TMA 50
3.4.3 Les prédicats non verbaux 61
3.4.4 La négation 66
3.4.5 Les articles 73
3.4.6 Les séries verbales 80
3.4.7 Conclusion 81

Cette partie présente les travaux que j'ai effectués sur la syntaxe des créoles santoméens et guadeloupéens. Elle vise à exposer la méthode, les présupposés et les outils que j'ai utilisés. Il ne s'agit pas d'une grammaire complète et je renvoie le lecteur vers les travaux cités pour les détails de l'analyse.

3.1 Quel est l'objet d'étude ?

3.1.1 Continuum ou diglossie ?

L'idée d'un continuum et la distinction entre basilecte, mésolecte et acrolecte date, pour les créoles au moins, des années 60 (Stewart (1969); DeCamp (1971)). Celle-ci opère un classement entre variétés selon leur proximité avec la langue lexificatrice (v. chapitre précédent). Dans leur souhait (louable de mon point de vue) de donner aux langues créoles le statut de langue 'autonome' par rapport à leur langue lexificatrice, certains créolistes ont théorisé le fait de n'accorder d'attention qu'à la forme basilectale. Cependant, ceci amène à la construction d'un objet parfois caricatural. [1] Un exemple particulièrement net qui illustre cette approche est Maurer (1995). En effet, on trouve dans cet ouvrage la volonté explicite d'écarter de la description de l'angolar tout ce qui s'approche trop du portugais et du créole principal de l'île (le forro). Bien que ses études reposent sur un travail de terrain et de corpus (Ph. Maurer met à disposition ses sources ce qui mérite d'être souligné !), il revendique la réécriture des parties qui ne sont pas de l'angolar basilectal.

> "L'objet de notre étude étant la description de l'angolar et non pas les interférences du santoméen et du portugais sur l'angolar, nous avons éliminé, avec l'aide de nos informateurs, les emprunts à ces deux langues là où l'angolar possède ses propres moyens d'expression." (p.158)

On comprend bien que la construction de la langue angolare (du 'vrai' créole) pose un problème majeur d'adéquation des analyses par rapport à la réalité constatée.

Face à la variation et au "mélange des langues", deux approches dominent :

— celle du continuum identifiable (v. plus haut),

— celle de la diglossie.

1. Voir à ce propos les excellents arguments de Mufwene (2001a).

Tabouret-Keller (1988) décrit très bien les enjeux et les conceptions à l'oeuvre dans ce débat.

On trouve une illustration de ce débat dans Zribi-Hertz (2011) et dans les travaux qu'elle conduit sur le haïtien notamment.

Il me semble, comme je l'avais esquissé dans ma thèse, qu'il est possible de prendre le problème par un autre bout et de partir des données recueillies pour construire dans un second temps un modèle théorique couvrant les données.

3.1.2 Corpus de données de terrain

Dans ce contexte, Jean-Louis Rougé et moi avons travaillé sur le recueil de données 'naturelles', c'est-à-dire non retouchées par le linguiste. Nous avons donc recueilli des enregistrements (souvent des récits de vie, quelques contes néanmoins) à São Tomé, à la fois auprès de locuteurs du forro (le créole majoritaire de l'île), des Angolares (sud de l'île) et d'une vieille dame Tonga (descendants de contractuels vivant dans les plantations de l'île), ce qui est décrit dans Rougé and Schang (2012). Nous avons fait de même en Guyane avec des locuteurs du créole guyanais standard, du français et du portugais. Ces enregistrements sont en partie accessibles sur le site de Cocoon (Huma-Num) https://cocoon.huma-num.fr/exist/crdo/search2.xql, depuis 2006 pour certains. Ces enregistrements sont disponibles gratuitement et sans restriction d'accès, permettant aux linguistes d'accèder directement aux données sur lesquelles nos études s'appuyent.

Cela m'a conduit également à proposer, en collaboration avec Anne Zribi-Hertz, un projet à la fédération TUL visant à recueillir des enregistrements en créoles martiniquais et guadeloupéens. Ce corpus a été réalisé et est accessible sur le site de Cocoon (CRDO) et il est identifié comme Glaude (2013). Les transcriptions ont été reprises par des étudiants de l'université d'Orléans (non diffusés) et ont servi comme ressources pour mes travaux sur le guadeloupéen.

Cette ressource a été utilisée par plusieurs collègues pour des projets de TAL, Millour and Fort (2018) et Schang, Antoine, and Lefeuvre-Halftermeyer (2017)

par exemple, mais aussi par des membres de SEEPiCLa à des fins de recherche (morphologie notamment).

D'autres projets sont en cours. Notamment la valorisation de 25 heures de kriol (Guinée-Bissau) recueillies par Jean-Louis Rougé. Ceci fait l'objet d'un projet en cours en partenariat avec les universités de Ziguinchor, Dakar et Coimbra. Ce projet a fait l'objet de communications (TALAf 2018 notamment, v. Mangeot and Schang (2018)). La méthologie est la suivante :

— environ 6 heures d'enregistrements ont été transcrits par un locuteur natif,
— nous utilisons le logiciel LIG-AIKUMA (Gauthier et al. (2016)) afin de procéder à du respeaking (répétition de segments) dans le but d'obtenir du signal de bonne qualité en parallèle du signal brut.
— nous allons ensuite utiliser le logiciel Persephone, basé sur TensorFlow, (Adams et al. (2018)) afin de réaliser une transcription phonémique du segment répété.

Enfin, pour le guadeloupéen, Juliette Sainton (ÉSPE de Guadeloupe) a saisi la perche tendue lors de ma communication au colloque des 40 ans de la Faculté de Linguistique Appliquée de Port-au-Prince et participe activement au développement de la grammaire électronique du guadeloupéen en fournissant un corpus de phrases.

3.2 Grammaires électroniques

Pour quoi faire ?

Je me suis trouvé, dans mes premières années de linguiste, face à des créolistes contestant mes exemples en forro car Untel [2] avait écrit un article qui ne donnait pas ce type d'exemples. Cette question de crédibilité, lorsque l'on est jeune docteur et moins à l'aise en anglais que d'autres, est essentielle. J'ai donc commencé

2. Le travail sur les langues peu connues autorise certains linguistes à des libertés parfois surprenantes avec la réalité. Ainsi, un linguiste allemand (dont l'article a été souvent mentionné à l'appui de thèses morphosyntaxiques), utilise des exemples en santoméen mal traduits, et place les créoles du Golfe de Guinée sur les mauvaises îles…

par mettre les fichiers son de mes exemples dans mes communications, et à chercher à adopter des modèles théoriques les plus transparents possibles.

La grammaire d'arbres adjoints (v. Abeillé and Rambow (2000) Abeillé (2002) Candito (1999) pour le français) m'a semblé être un modèle très intéressant car il possède la qualité d'être bien défini avec deux opérations seulement (substitution et adjonction). Il a cependant le désavantage de reposer sur une communauté d'utilisateurs plus réduite que d'autres modèles tels que LFG (Lexical-Functional Grammar, v. Kaplan, Bresnan et al. (1982)) ou HPSG (Head-driven Phrase Structure Grammar, v. Pollard and Sag (1994) et Henri (2010), Hassamal (2017) pour le créole mauricien).

Au delà de l'implémentation pour la validation scientifique, il me paraît important de contribuer également au travail d'équipement des langues dites 'peu-dotées' en proposant des analyses qui peuvent servir en ingénierie des langues.

Cela rejoint également une préoccupation issue de ma formation d'africaniste : l'approche déclarative. Ceci mérite quelques éclaircissements.

Approche déclarative

Dans mes premiers travaux de terrains effectués en Côte d'Ivoire, j'ai eu recours à des questionnaires d'inventaire grammatical. Parmi ceux-ci se trouvaient principalement Bouquiaux and Thomas (1976) et un questionnaire dactilographié dont je ne connais pas l'origine et qui m'avais été donné par mon professeur encadrant.

A la différence des questionnaires d'orientation générativiste, ces questionnaires sont conçus pour établir rapidement une grammaire ayant une couverture large des phénomènes, mais ils ne permettent pas de cerner dans la profondeur un phénomène. Par exemple, la grammaire générative propose de nombreux tests (v. Radford (1988, 1997) entre autres) compatibles avec les analyses habituelles de la grammaire générative. Toutefois, les analyses produites par ces tests, si elles sont précises sur un domaine, ont le désavantage d'être trop catégoriques, en particulier lorsque l'on travaille sur des langues très peu décrites, pour lesquelles le

moins que l'on puisse faire, c'est d'être très prudent.

En revanche, les questionnaires du type Bouquiaux and Thomas (1976) permettent de dire "j'ai trouvé ceci dans telle langue" et fonctionnent comme un inventaire, mais pas comme une analyse complète en lien avec une grammaire universelle. Dès lors qu'on se donne la liberté de ne pas (ou pas complètement) suivre les analyses proposées dans le cadre génératif, ces inventaires sont intéressants.

Paradoxalement, déclarer une structure arborescente en TAG (un arbre élémentaire) est une tâche plus modeste. Elle consiste à dire : "j'ai trouvé cette forme dans mon inventaire et je propose cet arbre. Il se trouve que cela permet de dériver les phrases attendues". Cela ne dit que peu de choses des structures de la langue interne, ce qui est une autre tâche. De ce fait, les analyses que j'ai proposées en TAG sont largement compatibles avec les analyses de Creissels (1994, 1991, 2006, 1995), bien plus qu'avec certaines analyses du programme minimaliste.

J'admets sans aucun problème qu'il y a *des* grammaires TAG pour un ensemble de phrases donné, certaines meilleures que d'autres assurément, mais rarement *une seule* grammaire TAG pour un problème.

Cette approche me permet d'ailleurs de proposer plusieurs codages alternatifs d'un même phénomène, en évaluant alors les conséquences de ces codages. J'ai bien conscience que pour certains collègues qui ne travaillent pas dans un tel modèle, renoncer à des explications éclairantes pour la Grammaire Universelle constitue une forme de recul théorique. Ce n'est cependant pas mon sentiment.

TAG

J'ai donc utilisé le modèle TAG à partir de mes travaux de thèse (Schang (2002)). J'ai proposé dans ceux-ci des fragments d'une grammaire TAG du saotomense qui n'ont pas fait l'objet d'une implémentation. Il aura fallu attendre la rencontre de mes collègues orléanais du LIFO et la découverte de leurs travaux (aux alentours de 2011) pour que je reprenne les travaux en TAG et cherche à implémenter mes analyses. Ceci constitue le but que je me suis toujours fixé. L'im-

plémentation constitue pour moi la preuve que l'analyse est correcte et au delà, permet de découvrir des problèmes qui n'apparaissent pas sur papier (au moins au premier regard). Cet aller-retour entre les données et l'implémentation constitue l'essentiel de ma démarche. C'est également ce que j'enseigne en L3 (cours sur les Grammaires d'Unification) dans le cadre HPSG (avec LKB) ou LFG (avec XLFG). C'est également aussi cette méthodologie que j'ai proposée pour la thèse que j'ai co-encadrée (voir Magnana Ekoukou (2015)). Elle correspond bien à la méthode d'analyse que j'utilise depuis mes débuts en linguistique et qui repose en grande partie sur les ouvrages de Denis Creissels (Creissels (1995, 2006)) .

Parmi les différents modèles que j'ai pu enseigner et donc utiliser, TAG est le modèle qui me semble correspondre le mieux aux besoins de la linguistique de terrain. Paradoxalement, c'est probalement le modèle qui est le moins utilisé pour la description de langues "rares", LFG et HPSG ayant une communauté d'utilisateurs plus vaste semble-t-il.

Metagrammaire

Il est bien connu que les grammaires TAG présentent le désavantage d'être 'verbeuses' dans le sens où elles demandent d'écrire un grand nombre d'arbres élémentaires. Elles sont également, c'est lié, assez coûteuses (plusieurs hommes.années pour les projets comme FTAG ou XTAG). L'utilisation d'une métagrammaire s'impose donc, surtout lorsqu'on est seul (ou presque) à écrire cette grammaire TAG.

Je vais décrire ici rapidement ce qu'est une métagrammaire avant de passer à un exemple complet : la métagrammaire du créole guadeloupéen.

Comme nous l'écrivions dans Duchier et al. (2017) :

> Parmi les approches semi-automatiques de création de ressources linguistiques, on peut distinguer les approches basées sur des langages de description des autres approches, par le fait qu'elles ne nécessitent aucune ressource externe préalable, et sont de ce fait adaptées à la création rapide de prototypes de ressources linguistiques *from*

> *scratch*. Les langages de description reposent sur des mécanismes d'abstraction permettant de définir de manière déclarative les régularités (voire dans certains cas les irrégularités) d'une langue. Cette description déclarative est ensuite traitée automatiquement (compilée) pour produire les unités décrites (par exemple les entrées d'un lexique ou d'une grammaire électronique). Comme nous le verrons, ces descriptions offrent divers avantages dont principalement (i) le fait de constituer en elle-même, de par leur structure modulaire et hiérarchique, une documentation des unités de la langue (mots ou règles syntaxiques dans notre cas), et (ii) le fait de permettre de vérifier la portée d'une théorie linguistique (en observant l'adéquation entre structures décrites et structures observées chez les locuteurs de la langue considérée).

XMG2 repose sur un langage de description qui nous permet de construire nos ressources.

On cite habiuellement Candito (1999) comme étant le premier travail totalement abouti sur les métagrammaires. D'autres travaux ont depuis suivi des pistes alternatives, parmi ceux-ci on trouve : Xia (2001), de La Clergerie (2010) et surtout, pour l'approche qui me concerne, Crabbé (2005).

XMG2 est un développpement qui fait suite à la thèse de B. Crabbé. Le principal apport des travaux de Simon Petitjean (Petitjean (2014) pour ne citer que sa thèse) est d'apporter une conception modulaire des métagrammaires. XMG2 permet la compilation de divers langages de description, à partir du moment où ceux-ci ont été définis formellement. XMG2 propose un ensemble de *briques de langage* (v. Petitjean, Duchier, and Parmentier (2016)) qui peuvent être assemblé de manière déclarative. Ces briques correspondent à diverses dimensions linguistiques et théoriques (structures de traits, arbres, formules sémantiques 'plates', semantic frames…).

Dans Duchier et al. (2017), nous présentons l'architecture de XMG2 sous la forme du schéma suivant en figure 3.1.

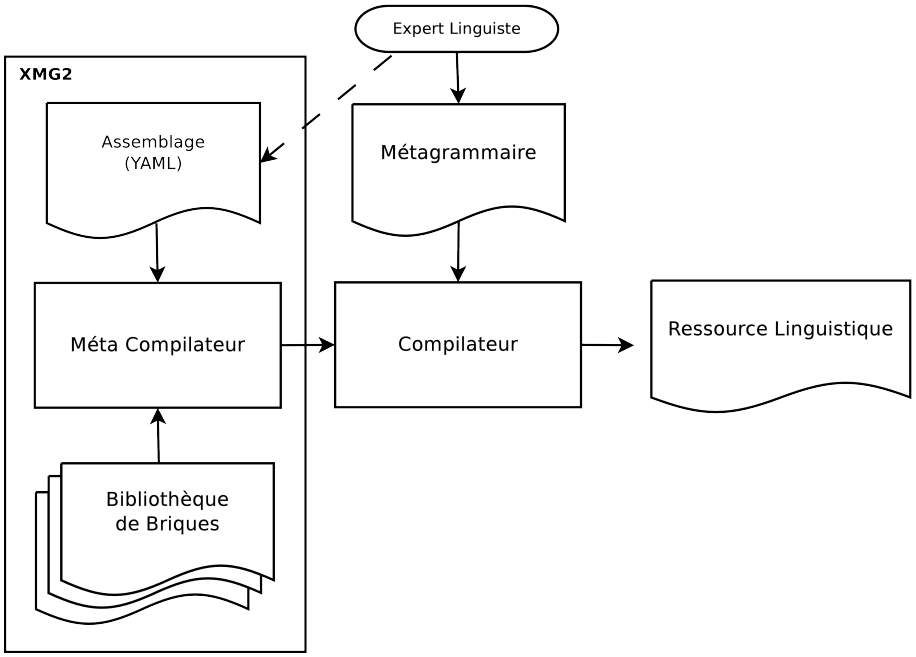


FIGURE 3.1 – Utilisation du système XMG2

Cette modularité a été utilisée pour sortir du cadre des arbres TAG et travailler sur la morphologie d'une langue peu décrite : l'ikota (Gabon, B25). Ce travail a été l'objet de la thèse de Brunelle Magnana-Ekoukou que j'ai co-encadrée avec Jean-Louis Rougé (Magnana Ekoukou (2015)). Dans ce travail ainsi que dans Duchier et al. (2012b), une morphologie à base de blocs (classes) permet de générer l'ensemble de formes fléchies d'un verbe ikota notamment.

3.3 TAG pour la créolistique : motivations

Dans les situations dans lesquelles sont parlés les créoles, la question du bornage des phénomènes que l'on décrit est cruciale. En effet, l'une des difficultés majeures dès lors que l'on souhaite décrire des faits attestés est de savoir si l'on a affaire à du créole ou pas.

(1) jénéralman de tout fason lè sé anivèsè a an timoun toujou ni anlo kado.
'Généralement, de toutes façons, quand c'est l'anniversaire d'un enfant, il y a toujours beaucoup de cadeaux.

On le voit bien en (1), il est difficile de dire s'il s'agit de créole ou de français sur les premiers mots (*jénéralman de tout fason*). La tentation, à laquelle je n'ai pas résisté, est de se focaliser sur la description du créole et de laisser de côté ce qui ressemble beaucoup au français. Toutefois, il convient, au minimum, de donner une méthode pour traiter ces cas.

Pour décrire ce qui se passe en (1), on peut imaginer que l'on dispose, comme c'est le cas pour les locuteurs du créole, d'une grammaire du français dans laquelle on peut aller piocher des arbres élémentaires. Ainsi, *jénéralman* et *de tout fason* peuvent être pris dans le stock d'arbres d'une grammaire TAG du français. La combinaison des arbres des deux langues peut poser des problèmes lors de l'unification mais la plupart du temps, les arbres fonctionnent de manière "étanche". Par exemple, dans les corpus dont je dispose, il n'y a aucun cas attesté de problème entre les traits fonctionnels (TMA et pluriel notamment). On ne trouve jamais des formes telles que :

(2) a. *nou ka mangeons
1PL IPFV manger.PRST.1SG
'Nous mangeons'

b. *sé les maisons lasa
PL DEF.PL maison DEM
'ces maisons'

Les arbres élémentaires me semblent être le niveau adéquat pour la description des interférences.

Ceci reste pour l'instant au niveau des hypothèses de travail[3], mais je considère qu'il s'agit d'un point théorique majeur à développer, notamment en lien avec la prosodie (v. plus loin). Veenstra (2009); López, Alexiadou, and Veenstra (2017)

3. Ceci figurait dans les objectifs d'un projet ANR-DfG qui n'a pas été retenu du côté français.

proposent des pistes intéressantes sur le lien entre le code-switching et la théorie des phases de la grammaire générative. Si ces travaux se confirment empiriquement, alors les TAG seraient une manière intéressante de formaliser ces questions d'alternance codique au sein d'un modèle génératif. En effet, Frank (2006) fait le lien entre le modèle TAG et la théorie des phases (v. Citko (2014) notamment) et souligne la proximité des modèles.

3.4 Contribution

Outre la constitution d'une ressource librement disponible et réutilisable, ce travail de réalisation d'une méta-grammaire apporte une contribution à la description linguistique du créole. Je présente dans les paragraphes qui suivent les apports principaux :

— traitement des marques de Temps, Mode et Aspect,
— les prédicats non-verbaux,
— la négation,
— les articles,
— les 'séries verbales'.

Ces quelques points seront l'occasion de présenter le fonctionnement de la métagrammaire (donnée en annexe) et les analyses linguistiques qui la guident. La section suivante présente rapidement les bases de la métagrammaire.

3.4.1 Utiliser une métagrammaire

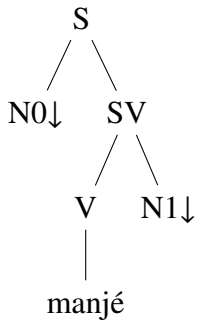
Décrire la structure argumentale des verbes constitue certainement l'utilisation la plus évidente de la métagrammaire. En effet, XMG permet de décrire les familles d'arbres élémentaires (ensemble d'arbres élémentaires associés à un prédicat donné et aux prédicats ayant le même cadre de sous-catégorisation) sans avoir besoin de recourir à des règles lexicales [4].

4. V. Abeillé (1993) pour une description des règles lexicales du français.

Partons d'un cas simple :

(3) Jan manjé diri
Jean manger riz
'Jean a mangé du riz'

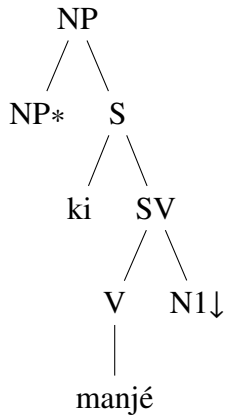
Pour décrire l'arbre élémentaire qui correspond au prédicat 'manger', on peut déclarer un arbre élementaire de la forme suivante :



A chaque nœud feuille correspond un argument placé ici dans une position dite 'canonique'. Les même arguments se retrouvent dans une autre configuration dans (4).

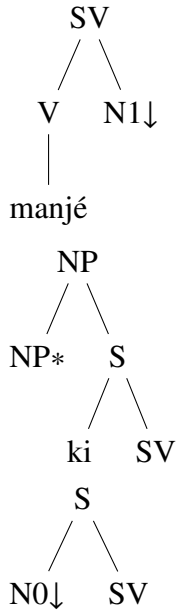
(4) Moun ki manjé diri vini
personne qui manger riz venir
'La personne qui a mangé du riz est venue'

L'argument sujet de 'manje' est ici représenté dans une autre position ; l'arbre élémentaire associé à 'manjé' dans cet exemple a donc la forme suivante :



On remarquera que l'on peut obtenir les deux arbres en factorisant le sous-arbre qui figure sous le SV.

On peut donc définir trois fragments qui vont se combiner :



Nous appellerons ces trois fragments respectivement : `Trans`, `RelSubj` et `CanSubj`.

On peut donc représenter les deux arbres bien formés des exemples (3) et (4) sous la forme d'une conjonction :

— `Trans` et `CanSubj` forment l'arbre en (3),

— `Trans` et `RelSubj` forment l'arbre en (4) ;

On peut donc exprimer la (mini-) famille `n0Vn1` (qui correspond aux verbes transitifs de notre petit exemple) sous la forme suivante :

$\{\{CanSubj \vee relSubj\} \wedge Trans\}$

On le voit, on peut exprimer une extraction élégamment sans recourir à une transformation ou un déplacement, mais par une disjonction.

C'est ce mécanisme qui sera utilisé par la suite dans cette grammaire du guadeloupéen.

3.4.2 TMA

Les analyses présentées ici reposent sur les travaux suivants : Schang et al. (2012b); Schang (2013, 2018). On notera que Schang et al. (2012b) porte sur le créole de São Tomé et ouvre la voie à un traitement similaire pour le guadeloupéen.

Les marqueurs TMA ont fait l'objet de l'attention des créolistes depuis toujours. Schuchardt en fait la description dans ses écrits et Bickerton leur accorde une place centrale dans sa théorie. Winford (2012) présente une synthèse et un état de l'art assez précis des enjeux autour de ces marqueurs. La question essentielle concerne le statut de ces élements. Pour le guadeloupéen, il s'agit de *ka* (habituel et imperfectif), *ké* (prospectif), *té* (antérieur) [5, 6].

On peut avoir une idée des valeurs des marques du guadeloupéen dans le tableau suivant :

Dans Schang (2000), les marqueurs TMA du forro sont les ancres d'arbres élémentaires avec un noeud pied V* permettant de les adjoindre sur le verbe. Des traits sur le noeud pied permettent de bloquer les combinaisons indésirables (**xka tava V* est bloqué mais *tava ka* est autorisé par exemple). Vaillant (2008a)

5. Les valeurs données ici sont celles habituellement données pour ces marques, v. Pfänder (2000); McCrindle (1999) notamment

6. Pour le santomense (forro), on trouve *ka* (imperfectif et habituel), *xka* (progressif), *tava* (antérieur).

TABLE 3.1 – Marqueurs TMA

VALEUR	FORME
Accompli	manjé
Inaccompli (Itératif etc.)	ka manjé
Futur (Prospectif)	ké manjé
Passé accompli	té manjé
Passé inaccompli	té ka manjé
Irréaliste passé	té ké manjé
Irréaliste inaccompli	té ké ka manjé

adopte cette même représentation pour le martiniquais et le guadeloupéen. Cette façon de procéder permet de dériver le bon ensemble de phrases. Techniquement, cette solution fonctionne. Cependant, il m'a semblé nécessaire de revenir sur cette représentation des TMA pour des raisons linguistiques.

Dans Schang (2018) notamment, je défends l'idée que les TMA ne peuvent pas ancrer des arbres élémentaires pour les raisons suivantes

Comportement syntaxique des TMA

On ne trouve jamais les TMA en réponse à une question (**Oui, an ka !* vs *Yes, I did !*), ce qui les distingue des auxiliaires de l'anglais par exemple.

De plus, on ne peut pas les coordonner, à la différence des auxiliaires en français (*à celui qui a été ou sera condamné pour un vol qu'il a ou aura commis lui-même, ou qui a ou aura transigé…*).

(5) *Jan ka é ké manjé.
Jean ipfv et prosp manger
Attendu : 'Jean mange et mangera'

Si l'on veut obtenir une suite temporelle, la répétition du verbe est la norme :

(6) an fouté y, an ka fouté y, an ké fouté y !
'J'ai déjà gagné, je gagne cette fois-ci et je gagnerai encore ! (version polie,

le verbe 'fouté' ayant bien entendu un autre sens...)'

Le clivage du prédicat possible mais sans les TMA, seul l'élément lexical est clivé (à l'exception des modaux).

(7) sé monté nou ka monté pou nou rivé la nou ka alé.
c'est monter 1pl IPFV monter pour 1pl arriver là 1pl IPFV aller
'On va grimper, grimper pour arriver là où l'on va' (intensif)

(8) a. *sé **ka** monté nou ka monté...
b. *sé **té** monté nou té monté...

La négation est également exclue du clivage :

(9) Sé/pou (*pa) monté, nou pa monté
c'est/pour NEG monter 1pl IMPERF monter
'Nous ne sommes PAS monté.'

Mais les modaux *pé* et *vlé* peuvent être clivés avec la négation :

(10) Sé vlé pa, i vlé pa !
c'est vouloir NEG 3SG vouloir NEG
'Ah ça, il (ne) veut pas, il (ne) veut pas !'

On le voit donc, les TMA n'ont pas le comportement 'typique' attendu d'un élément autonome syntaxiquement. Il est donc questionnable de les placer en ancre d'un arbre élémentaire.

Mais il y a d'autres arguments encore, qui les rapprochent d'éléments assemblés en morphologie (si tant est que l'on souhaite distinguer nettement les opérations syntaxiques et les opérations morphologiques).

En effet, le marqueur *ké* se trouve fusionné avec la négation *pa*, ce qui donne *péké* et même *péé*.

(11) Jan péké manjé
Jean NEG-PROSP manger
'Jean ne mangera pas'

Ce type de comportement s'inscrit dans les phénomènes traités habituellement par la morphologie. Toutefois, on remarquera que les TMA n'ont pas le comportement attendu d'un préfixe car des adverbes peuvent s'intercaler entre eux.

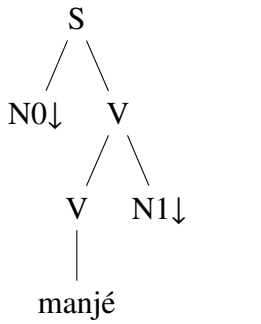
(12) Pyè te ja ka vin
Pierre ant. déjà ipfv venir
'Pierre était déjà en train de venir'

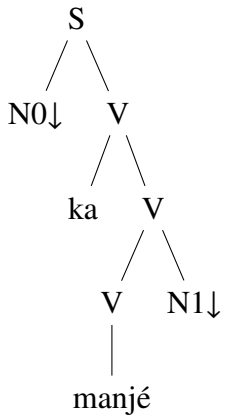
En conséquence, j'ai proposé de décrire ces TMA comme des co-têtes du prédicat (nominal ou verbal) en suivant l'idée proposée par Frank (2002) d'incorporer les projections fonctionnelles dans les arbres élémentaires. Cela me semble compatible avec les idées avancées généralement en grammaire générative (Grimshaw (2000, 1991) notamment) et tendrait vers les analyses de Frank (2006, 2002), en conservant toutefois une approche lexicaliste (au sens où la composition des mots n'est pas lisible directement en syntaxe).

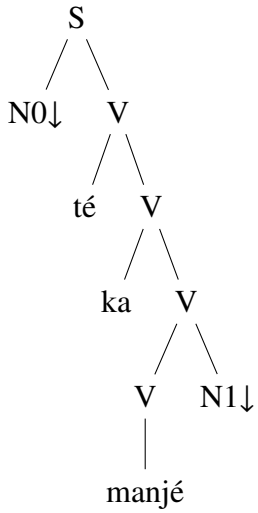
Par ailleurs, il est bien connu que le marqueur *ka* s'interprète différemment selon la classe de prédicat sur lequel il porte. Par exemple, *ka* sera à interpréter comme inaccompli dans *Jan ka manjé* 'Jean est en train de manger' mais comme un itératif dans *i ka bèl lè i pengné konsa* 'elle est belle quand elle est peignée comme cela'. On peut trouver ici un avantage à considérer que *ka* n'est pas interprétable seul, mais ne peut s'interpréter que dans la séquence ka+V. C'est-à-dire, en analysant un objet de la largeur de l'arbre élémentaire du prédicat.

Traitement des TMA dans la métagrammaire

Traiter les TMA comme des co-ancres du prédicat conduit à générer un arbre élémentaire pour chacune des formes 'conjuguées' du verbe. On trouvera donc pour les formes 'canoniques' de *manjé* les formes suivantes qui correspondent respectivement à 'Jean a mangé', 'Jean est en train de manger' et 'Jean était en train de manger' :

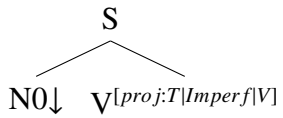




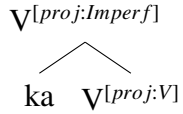


Je ne vais pas développer l'analyse pour toutes les formes, mais on le comprendra aisément, il est possible de factoriser des parties de l'arbre et d'utiliser les mécanismes de la métagrammaire pour assembler les fragments.

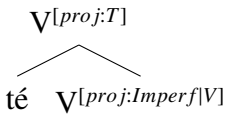
a.



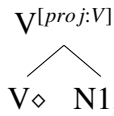
b.



c.



d.



Bien entendu, comme les TMA ont un ordre linéaire contraint, il est nécessaire d'utiliser un mécanisme filtrant les combinaisons correctes. J'utilise un trait `proj` (projection) qui va bloquer les combinaisons invalides (telles que **ka té V* par exemple). On appelera :

(a.) : `CanSubject`,

(b.) : `Imperf`,

(c.) : `Anterior`,

(d.) : `Transitive`

Le fait de disposer de noeuds distincts dans l'arbre, avec un niveau de projection donné, permet de réaliser l'adjonction des adverbes tels que *ja* 'déjà' au bon niveau. Ce dernier aura un nœud pied portant les traits `[proj:Imperf|V]`.

On peut donc facilement représenter le fonctionnement des TMA en guadeloupéen sous la forme de conjonctions et de disjonctions. J'ai choisi de faire appel à un artefact pour représenter cela : une classe `none` qui n'a aucun contenu [7].

7. Il ne s'agit pas d'une position vide ou encore d'un morphème non prononcé, mais simplement d'une facilité d'écriture.

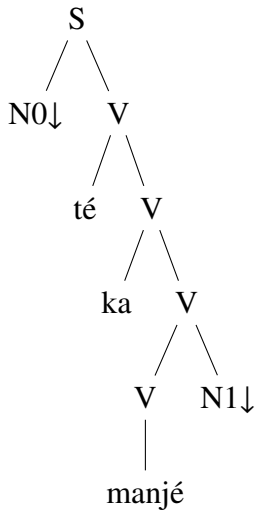
On trouvera donc :

```
{ { Prosp(ective) (ké) | None };
{ Imperf(ective) (ka) | None };
{ Anterior (té) | None };
Transitive}
```

On peut développer de la manière suivante : on a {(`Prospective` ou `None`) & (`Imperfective` ou `None`) & (`Anterior` ou `None`) & `Transitive`}.

Revenons à l'exemple de départ. L'arbre (13) est constitué de la conjonction des classes `CanSubject`, `Anterior`, `Imperf` et `Transitive`.

(13)



Intégrer les TMA comme co-ancre des verbes (et autres prédicats) s'avère donc à la fois possible (la grammaire obtenue est faiblement équivalente à une grammaire pour laquelle les TMA ancreraient des arbres élémentaires auxiliaires et surtout souhaitable. En effet, les questions de localité (à quel niveau doivent s'interprêter les éléments syntaxiques) me paraissent recevoir un traitement adéquat :

— les TMA ne sont pas interprétables sans référence à la classe aspectuelle du prédicat qu'ils accompagnent ;

— ce ne sont pas des affixes, des positions syntaxiques sont disponibles pour l'adjonction (adjoining) des modifieurs adverbaux ;

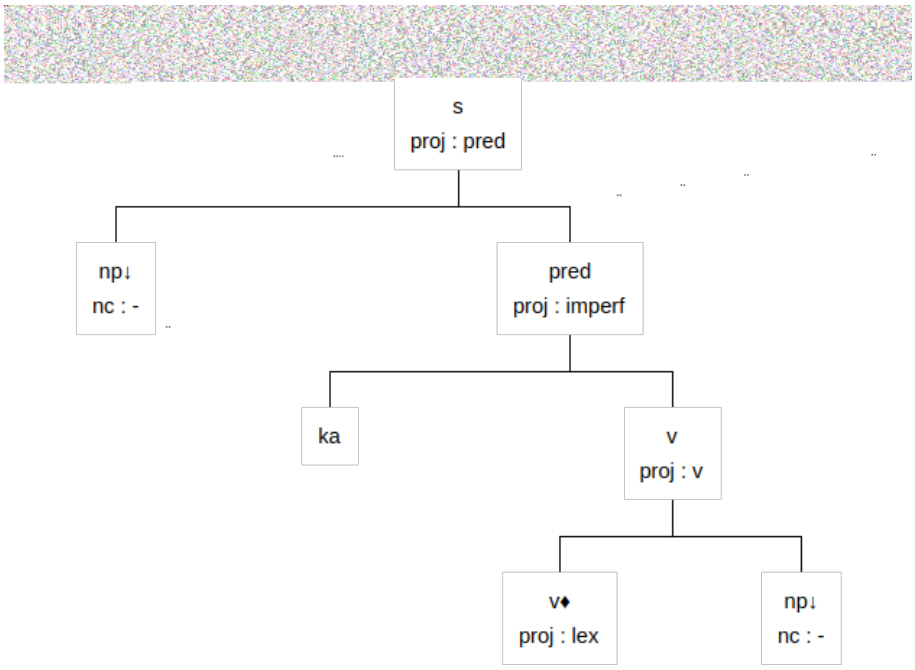
— les arbres de dérivations peuvent aisément servir de pivot pour une traduction français/créole, créole/français. Prenons par exemple *Jean mangeait* et sa traduction *Jan té ka manjé*, on aura pour les deux langues la dérivation suivante :

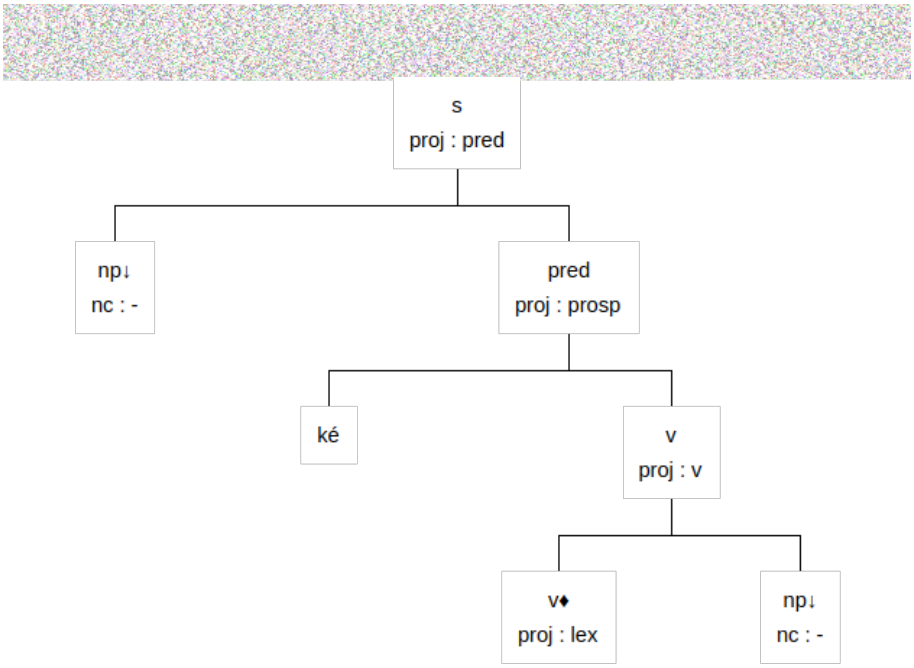
α1 manjé {past ;ipfv} α1 manger {past ;ipfv}

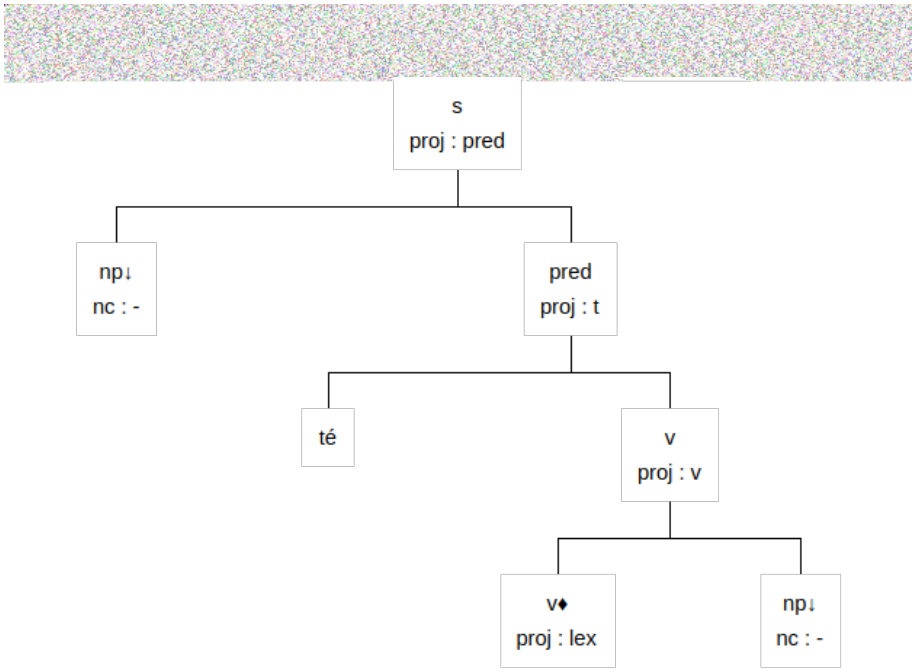
α2 Jan α2 Jean

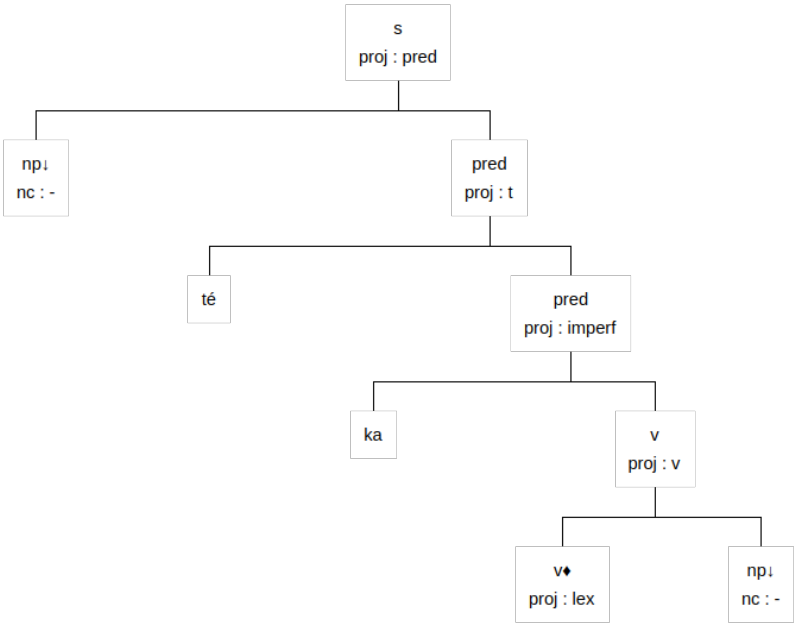
— enfin, cette analyse permet de poser sous un autre angle la question de l'absence de morphologie dans les langues créoles. Bien entendu, celle-ci ne permet pas d'affirmer que le guadeloupéen possède une morphologie flexionnelle très riche, cependant, celle-ci n'est pas inexistante.

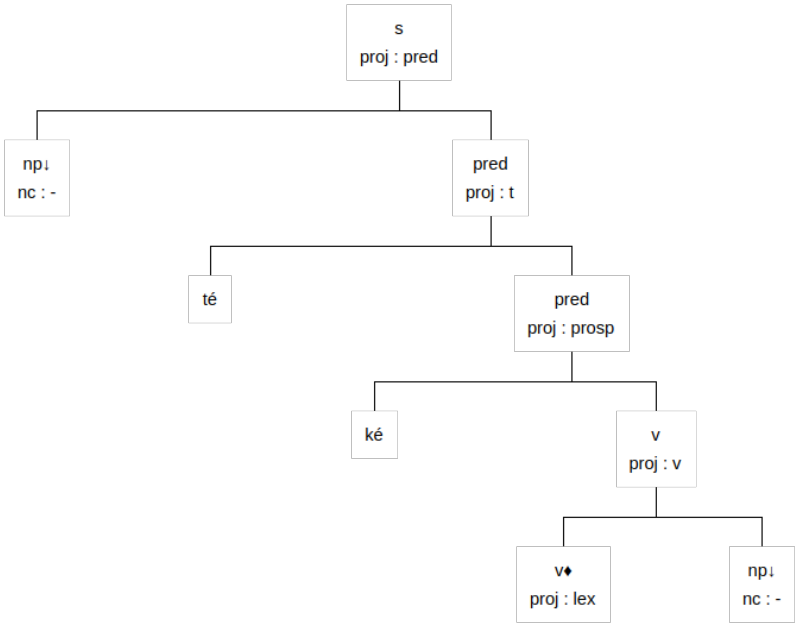
Voici les principaux arbres élémentaires incluant des TMA :




s
proj : pred
np↓
nc : -
pred
proj : prosp
ké
v
proj : v
v♦
proj : lex
np↓
nc : -


s
proj : pred
np↓
nc : -
pred
proj : t
té
v
proj : v
v♦
proj : lex
np↓
nc : -


s
proj : pred
np↓
nc : -
pred
proj : t
té
pred
proj : imperf
ka
v
proj : v
v♦
proj : lex
np↓
nc : -



3.4.3 Les prédicats non verbaux

Le créole offre la possibilité de créer des phrases complètes autonomones composées d'un sujet et d'un prédicat non-verbal (voir (14) et (15)).

(14) a. moun la doktè/bèl/gwo/malad
personne DEF docteur/beau/gros/malade
'Cette personne est docteur/belle/grosse/malade' .'

b. nou Jeanny
1PL Jeanny
'On est Jeanny. (on vote pour Jeanny)'

c. timoun la pli gwan ki Jan
enfant DEF plus grand que Jean
'Cet enfant est plus grand que Jean'.

(15) a. Jan anba
Jean en-bas
'Jean est en bas'

b. Jan adan lanméri
Jean dans mairie
'Jean est dans la mairie.'

c. lékol la lwen
école DEF loin
'L'école est loin'

Comme proposé dans Vaillant (2008b), je considère qu'il s'agit de prédicats non-verbaux et qu'il n'existe pas de copule, même non prononcée (v. Henri and Abeillé (2007) pour une analyse similaire en mauricien).

Il est assez simple de construire ces phrases contenant un prédicat nominal ou adjectival avec XMG, en réutilisant les fragments déjà utilisés pour les prédicats verbaux :

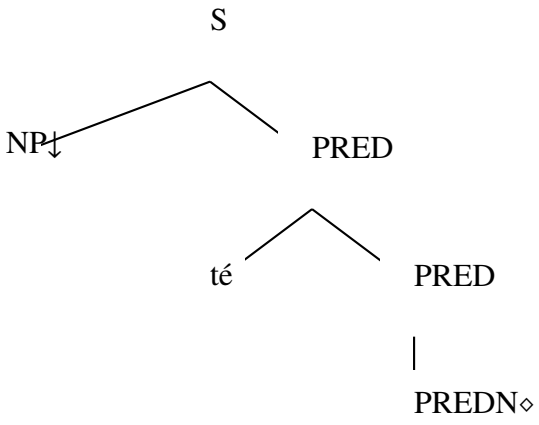
```
class NonVpredn
export ?X ?Y
declare ?X ?Y
{ <syn>{
        node ?X (color=white)[cat = pred, proj = predn]{
                node ?Y (mark=anchor,color=black)[cat= predn,proj=lex]
                }
        }
}
class NonVpreda
export ?X ?Y
declare ?X ?Y
{ <syn>{
        node ?X (color=white)[cat = pred, proj = preda]{
                node ?Y (mark=anchor,color=black)[cat= preda,proj=lex]
                }
```

```
        }
}
class zeroCop
{
CanSubject[];
{NonVpredn[] | NonVpreda[] } ;
TMA[]
}
```

On peut alors assembler ces fragments avec les TMA et un sujet au sein d'une classe `zeroCop`.

Ceci permet de produire des arbres, comme par exemple `n0téPREDN` ci dessous, qui correspond à l'exemple (16).

(16) Jan té doktè
Jean ANT docteur
'Jean était médecin'



Il y a plusieurs difficultés cependant :

— les prédicats nominaux ne peuvent pas contenir un article défini, indéfini, démonstratif (17). C'est pour cela que la classe `NonVpredn` contient un noeud qui n'autorise que la projection `[cat: predn,proj=lex]` qui va

bloquer la possibilité d'associer un fragment d'arbre contenant un article.

(17) a. Jan sé on doktè
Jean être un docteur
'Jean est un docteur'

b. *Jan on doktè

— Les extractions du prédicat nécessitent la présence de la copule (18) :

(18) a. ki koté nou yé ?
Q côté 1PL être
'Où sommes-nous ?'

b. pèsonn (ki) nou yé
personne (que) 1PL être
'la personne que nous sommes...'

c. Sé malad li yé.
être malade 3SG être
'Il est vraiment malade'

On peut donc représenter les contraintes sur ces constructions en assemblant un fragment `Yé` avec la négation (ou pas), des éléments extraits (`{RelObject[] | WhObject[]}`) mais pas de TMA :

```
 class Yé
export ?X ?Y
declare ?X ?Y
{ <syn>{
        node ?X (color=white)[cat = v, proj = v]{
                node ?Y (mark=anchor,color=black)[cat= beCopFin, proj = lex]
                }
        }
}
class Vyé
{
{RelObject[] | WhObject[]} ;
```

```
{Neg[] | None[]} ;
Yé[]
}
```

Un autre point mérite une attention particulière : les prédicats locatifs (19).

(19) a. Jan anba
Jean en-bas
'Jean est en bas'

b. Jan adan lanméri
Jean dans mairie
'Jean est dans la mairie.'

c. lékol la lwen
école DEF loin
'L'école est loin'

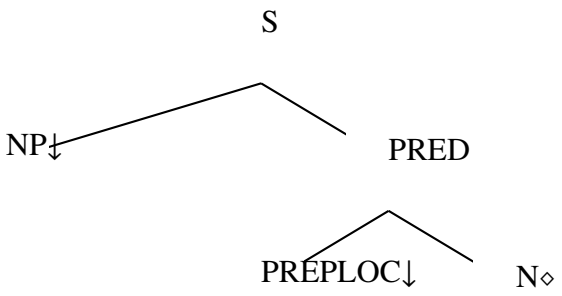
d. Jan laplaj
Jean plage
'Jean est à la plage'

e. *Jan plaj

On remarque que l'on va trouver différents éléments : des adverbes, des syntagmes prépositionnels, mais aussi certains noms sous une forme en la-N.

Les cas simples concernent les adverbes et les SP. Pour les premiers, il y a une classe `NonVpredadv` similaire aux classes déjà vues. Pour les SP, il faut créer une classe pour la préposition locative qui va pouvoir faire partie des prédicats non verbaux.

Les prédicats contenant une préposition locative ont donc la forme :



Le cas le plus surprenant est celui présenté en (19-d,e). En effet, certains noms ont deux formes : une forme locative (*laplaj*) et une forme 'normale' (*plaj*). Le trait (`nloc`) encodera cette différence dans le lexique. Je ne développerai pas plus ici, mais la différence entre les deux mots ne s'arrête pas là, elle concerne aussi la possibilité de s'associer aux articles et aux quantifieurs (v. Zribi-Hertz and Jean-Louis (2013) pour une analyse en terme de noms propres pour les formes en la-N).

Par ailleurs, ce codage dans la métagrammaire permet d'associer des sens différents [8] aux expressions en (20) étant donné la différence de structure.

(20) a. Jan lékol
Jean école
'Jean est à l'école'

b. Jan adan lékol
Jean dans école
'Jean est dans l'école'

3.4.4 La négation

L'analyse des éléments négatifs en guadeloupéen apporte d'autres arguments en faveur du traitement des TMA comme co-ancres du prédicat. Dans Petitjean and Schang (2018) nous détaillons le comportement des éléments négatifs en créole et proposons une analyse dans le cadre de cette métagrammaire.

8. Une sémantique à base de Frames permet de rendre compte assez bien de ce type de nuances, mais je n'ai pas encore développé cette dimension, autrement que par une preuve de concept sur une poignée d'exemples.

En particulier, nous montrons que les éléments négatifs ont des comportements hétérogènes. Comme l'écrivent V. Déprez et F. Henri dans l'introduction de Déprez and Henri (2018) à propos des données que nous présentons :

> The data clearly contradicts the idea of a creole prototype with a strict negative concord *à la Bickerton*. Guadeloupean clearly exemplifies a system with morphosyntactic expressions offering a wide range of interpretations, from concord, strict and non-strict, to double negation readings.

Avant de présenter les points clés de la métagrammaire concernant la négation, je me permets de présenter les données exposées dans l'APICS sur la négation en guadeloupéen.

La description annonce :

— une négation préverbale :

> Standard Negation before the verb.
>
> Example 50-191
>
> (21) Pyè pa vini.
> Peter neg come
> Peter did not come.
>
> Type : naturalistic spoken
>
> Source : Own fieldwork
>
> Colot and Ludwig (2013)

— seul *pa* est pris en compte dans la négation.

Bien que ces données ne soient pas entièrement fausses car cela correspond bien, grosso modo, à ce que l'on trouve en guadeloupéen, on ne peut pas en rester là. Nous venons de le voir, les modaux *pé* et *vlé* peuvent être suivis du marqueur négatif *pa*. Ceci mérite d'être signalé.

Ensuite, se contenter de *pa* comme marqueur négatif conduit à montrer que le guadeloupéen et le haïtien ont un fonctionnement semblable (visible sur les cartes

de l'APICS). Pourtant, lorsque l'on regarde en détail le fonctionnement des deux créoles, on constate des différences significatives, notamment sur les questions de localité, comme le soulignent Déprez and Henri (2018) dans la conclusion de l'ouvrage (portée à longue distance).

Losque l'on rentre dans les détails, on constate que :

— *pa* fusionne avec l'adverbe *anko* pour former *poko* et avec le prospectif *ké* pour former péké voire même péé :

(22) i poko vini
3sg neg.encore venir
'Il n'est pas encore venu'

(23) i péé vini
3sg neg.prosp venir
'Il ne viendra pas'

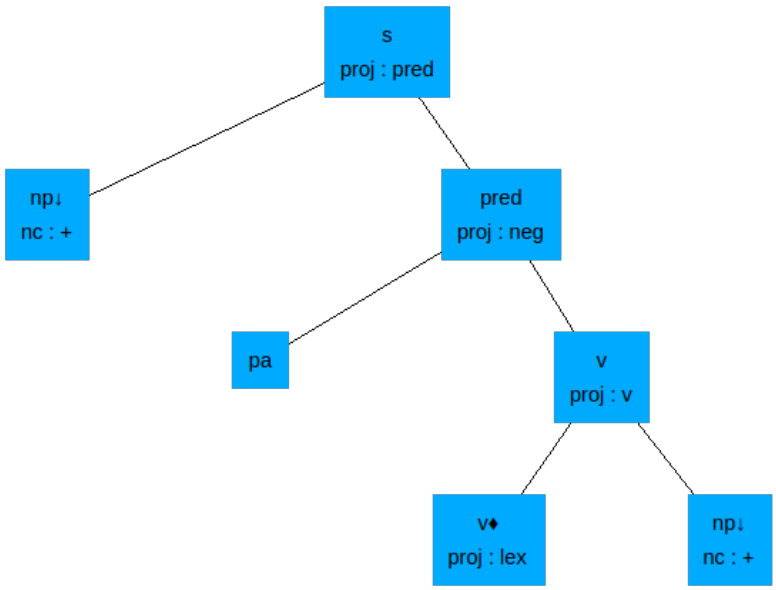
Dans la métagrammaire, il est donc nécessaire d'envisager plusieurs cas :

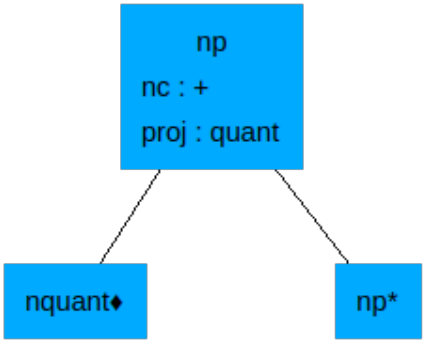
— les marques qui vont s'insérer au sein des TMA et déclencher la présence d'un trait `nc:+` sur les feuilles de l'arbre qui correspondent aux arguments du verbe (et plus généralement du prédicat). En l'absence de ces marques, les mots négatifs tels que *pon* 'aucun' ne peuvent pas s'unifier avec l'arbre du verbe.
— les autres marques qui ancrent un arbre élémentaire et qui demandent éventuellement la présence d'un marqueur négatif sur le prédicat.

C'est à dire que dans (24), la présence de *pa* déclenche l'attribution de la valeur `nc:+` sur le noeud auquel *pon moun* (nquant) va être associé. *pon* porte également un trait `nc:+` permettant l'unification.

(24) Jan pa vwè pon moun
Jean NEG voir aucun personne
'Jean n'a vu personne'

(25) *Jan vwè pon moun (même sens recherché)





L'adverbe *jen* 'jamais' qui a un comportement singulier mérite un traitement particulier (cf. Petitjean and Schang (2018) pour le détail de l'argumentation).

Lorsqu'il est sous sa forme emphatique, JANMÈ 'jamais' ne requiert pas la présence de *pa*, contrairement à la forme *jen*. *jen* est donc intégré dans la métagrammaire parmi les marqueurs préverbaux [9].

Le code XMG correspondant est donc :

```
class NegT
export ?X ?Y ?Z
declare ?X ?Y ?Z ?N

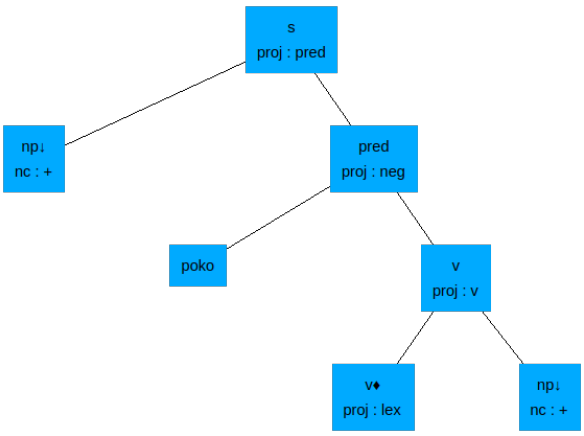
{ <syn>{
        node ?X (color=white)[cat = pred, proj = negT]{
                node ?Y (mark=flex,color=red)[cat= jen]
                node ?Z (color=black)[cat = @{pred,v},
proj = @{t,prosp,imperf,v,preploc,nloc,preda,predn}]
                }
        };
  Neg[]
}
```

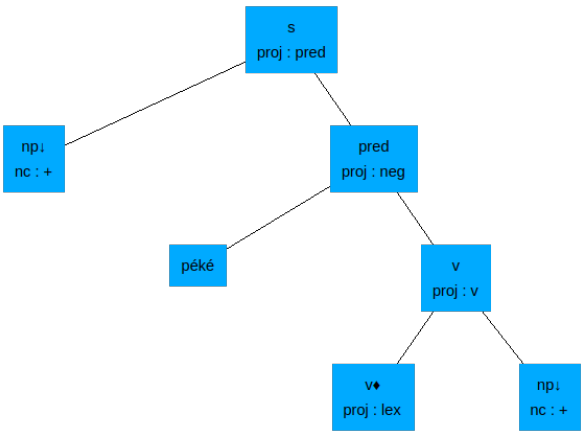
La co-présence obligatoire de la classe `Neg` permet d'exclure la présence de *jen* sans *pa*.

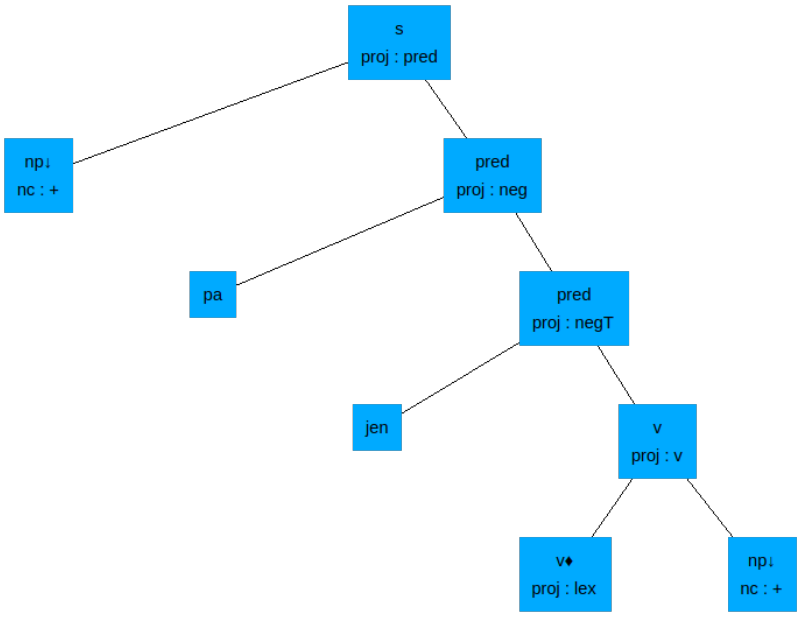
On trouve donc parmi les éléments négatifs préverbaux qui fonctionnement comme co-ancre du prédicat : *pa, poko, péké, jen*.

Ce qui donne donc les arbres suivants :

9. Marqueurs qui précèdent le prédicat pour être précis.


s
proj : pred
np↓
nc : +
pred
proj : neg
poko
v
proj : v
v♦
proj : lex
np↓
nc : +


s
proj : pred
np↓
nc : +
pred
proj : neg
péké
v
proj : v
v♦
proj : lex
np↓
nc : +



Je regroupe donc sous le label `TMA` (label commode mais impropre car la négation n'est pas à classer dans les TMA) l'ensemble de marques préverbales, négatives ou non, et leurs combinaisons :

```
class None

class TMA
{
  { Prosp[]  | None[] };
  { Imperf[] | None[] };
  { Tensed[] | None[] };
  {{ Neg[] | NegT[] | poko[] | péké[] }*=[nc= +] | None[]*=[nc= -] }
}
```

En d'autres mots, la classe `TMA` exprime l'ensemble des combinaisons licites

des marqueurs préverbaux. Le trait `nc:+` permet de décrire les phénomènes de concordance négative, comme cela est décrit dans Petitjean and Schang (2018).

Cette notation me paraît combiner assez bien les besoins de lisibilité du code et la puissance expressive de XMG.

3.4.5 Les articles

Je présente ici le traitement des projections fonctionnelles [10] associées au nom. En effet, il aurait été dommage de ne pas chercher à reprendre l'idée développée autour des TMA dans le domaine nominal.

La base du SN est constituée du nom nu (Bare N), entendu comme nom sans projection fonctionnelle 'explicite' [11]] :

np
|
n◇

En guadeloupéen, l'article défini *la* (plutôt associé à la valeur spécifique) occupe la position finale du NP.

(26) chyen/biten/moun/tab la
chien/chose/personne/table DEF
'Le chien, la chose, la personne, la table'

(27)

Il est donc nécessaire de bloquer toute adjonction qui se ferait au dessus de la projection de *la*. Pour cela, les adjonctions dans le NP, comme par exemple pour un adjectif épithète préposé (ci-dessous pour la classe XMG et les arbres dérivés de *Jan manjé bon diri* 'Jean a mangé du bon riz' et *Jan manjé bon diri la* 'Jean a mangé le bon riz'), se font sur le noeud N :

10. Au sens du paragraphe précédent.
11. Le nom nu est ici envisagé différemment de ce qui se fait dans certains travaux de grammaire générative, où il existe des projections fonctionnelles sans contenu phonologique.

```
class Adj %pour les épithètes, pied à droite
export ?X ?Y ?Z
declare ?X ?Y ?Z
{ <syn>{
        node ?X (color=black)[cat = n]{
                node ?Y (mark=anchor,color=black)[cat = adj]
                node ?Z (mark=foot,color=black)[cat = n]
                }
        }
}
```

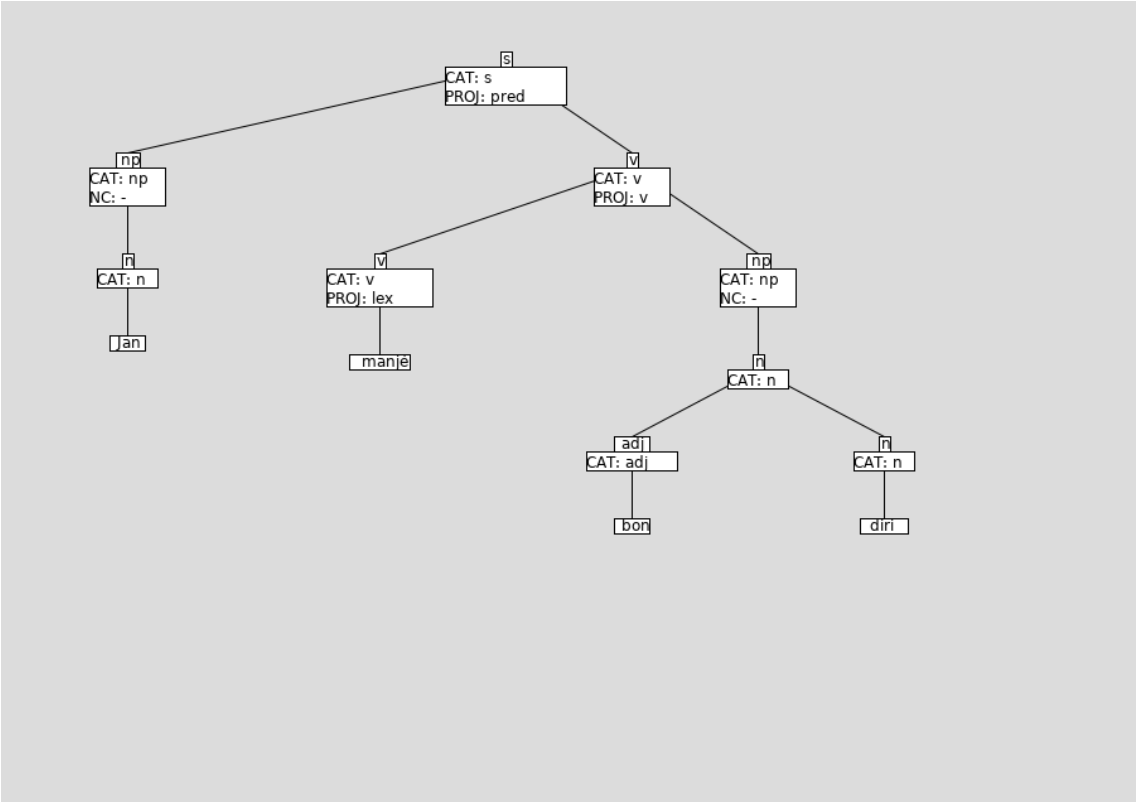


FIGURE 3.2 – Arbre dérivé de 'Jan manjé bon diri'

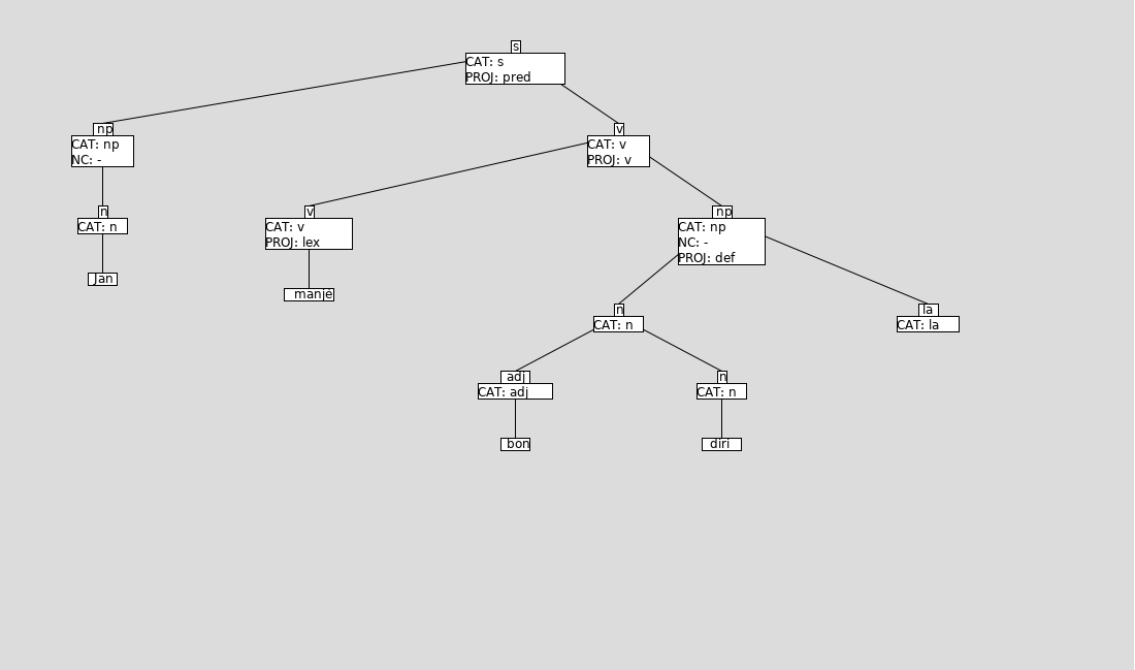


FIGURE 3.3 – Arbre dérivé de 'Jan manjé bon diri la'

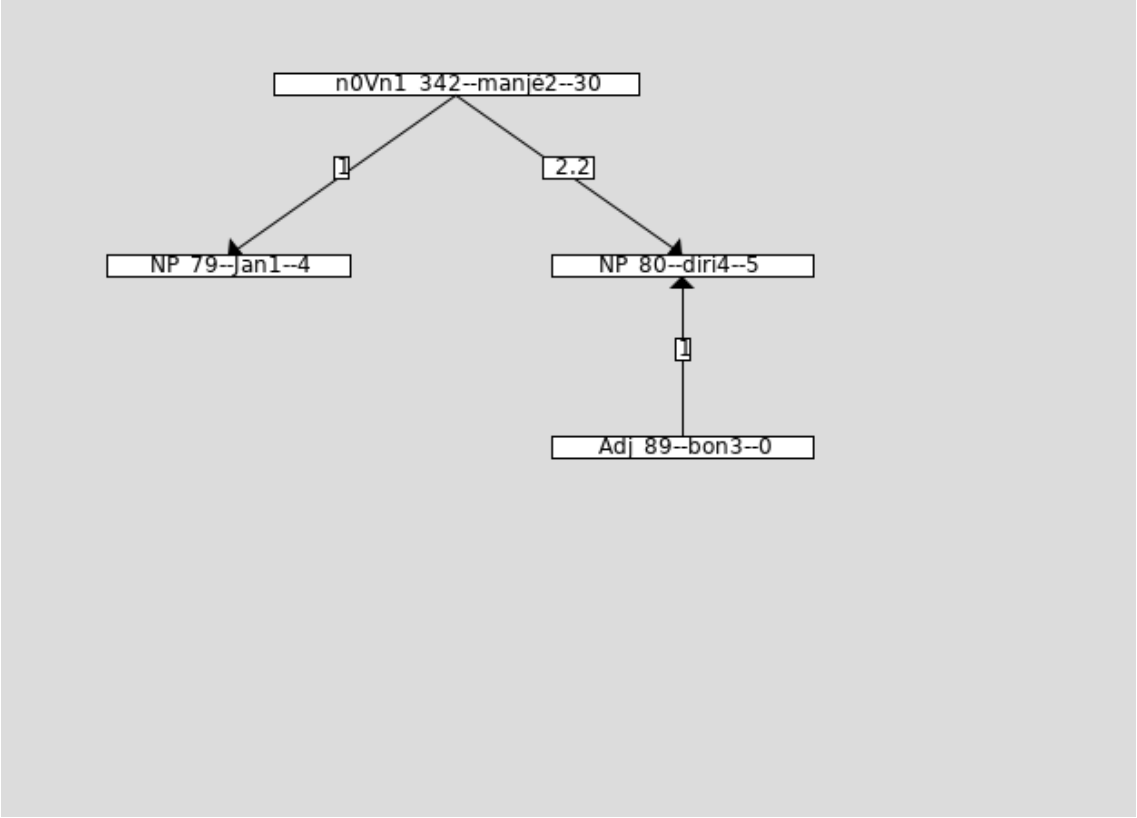


FIGURE 3.4 – Arbre de dérivation de 'Jan manjé bon diri la'

J'utilise le même principe des co-ancres fonctionnelles pour le pluriel (29) et le démonstratif (28) :

(28) Jan manjé bon diri lasa
Jean manger bon riz DEM
'Jean a mangé ce bon riz'

(29) a. Jan manjé sé bon diri lasa
Jean manger PL bon riz DEM
'Jean a mangé ces bons riz'

b. Jan manjé sé bon diri la
Jean manger PL bon riz DEF
'Jean a mangé les bons riz'

c. *Jan manjé sé bon diri
Jean manger PL bon riz
'Jean a mangé les/des bons riz'

En effet, les marqueurs *la, lasa* et *sé* se combinent comme les fragments TMA. En particulier, on remarquera que le pluriel *sé* demande que le nom soit défini, d'où la création d'une classe `DEF` qui rassemble de défini et le démonstratif.

```
class DefN
declare ?DF ?N
{
?DF = Defsat[];
?N = Noun[]
}

class DemN
declare ?DF ?N
{
?DF = Demsat[];
?N = Noun[]
}
```

```
class DEF
{
{Dem[] | Def[]}
}

class PlN
declare ?Pl ?N ?DF
{
?DF = DEF[];
?Pl = Pl[];
?N = Noun[]
}
```

En revanche, le marqueur indéfini *on, yon, an* 'un' est traité comme les autres quantifieurs (comme *kèk, anlo, anpil...* 'quelques, beaucoup'). Il ancre un arbre élémentaire auxiliaire qui s'adjoindra au nom.

```
class Quant
export ?X ?Y ?Z
declare ?X ?Y ?Z
{ <syn>{
        node ?X (color=black)[cat = np, proj = quant]{
                node ?Y (mark=anchor,color=black)[cat = quant]
 node ?Z (mark=foot,color=black)[cat= np]
                }
        }
}
```

Pour les 'low quantifiers' (comme *kèk* 'quelques') :

```
class lQuant
export ?X ?Y ?Z
declare ?X ?Y ?Z
{ <syn>{
        node ?X (color=black)[cat = n]{
                node ?Y (mark=anchor,color=black)[cat = lquant]
                node ?Z (mark=foot,color=black)[cat = n]
                }
        }
}
```

Les arbres dérivés suivants illustrent le traitement des quantifieurs :

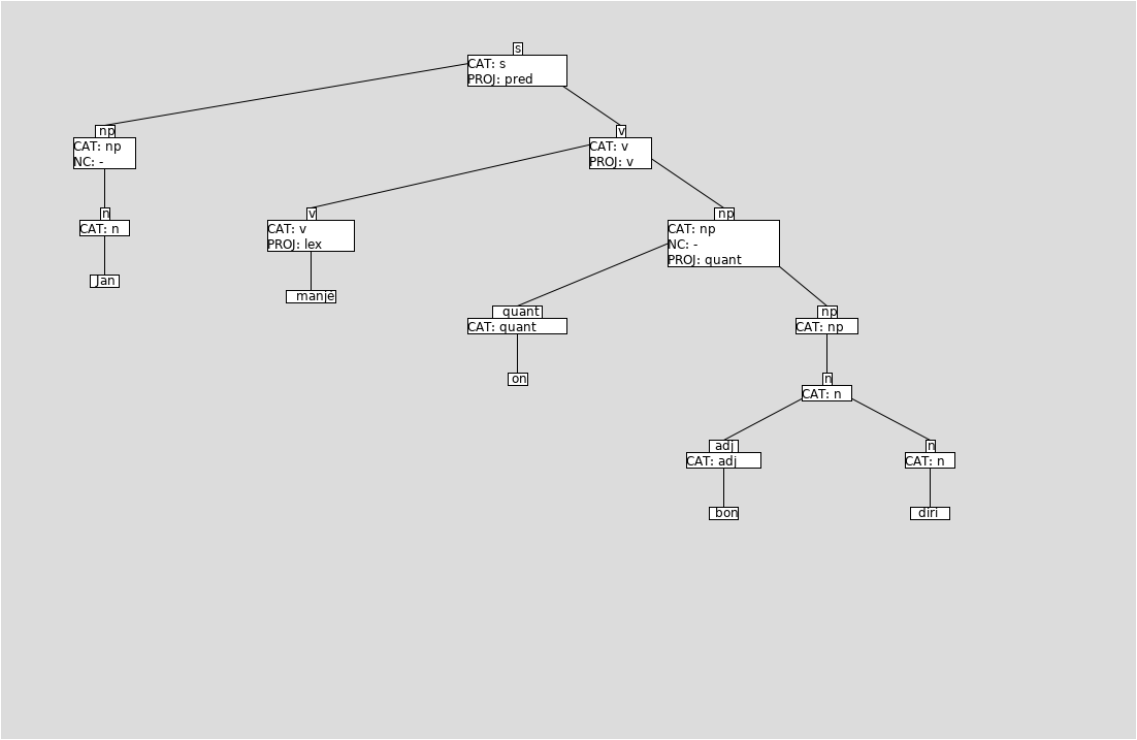


FIGURE 3.5 – Arbre dérivé de 'Jan manjé on bon diri'

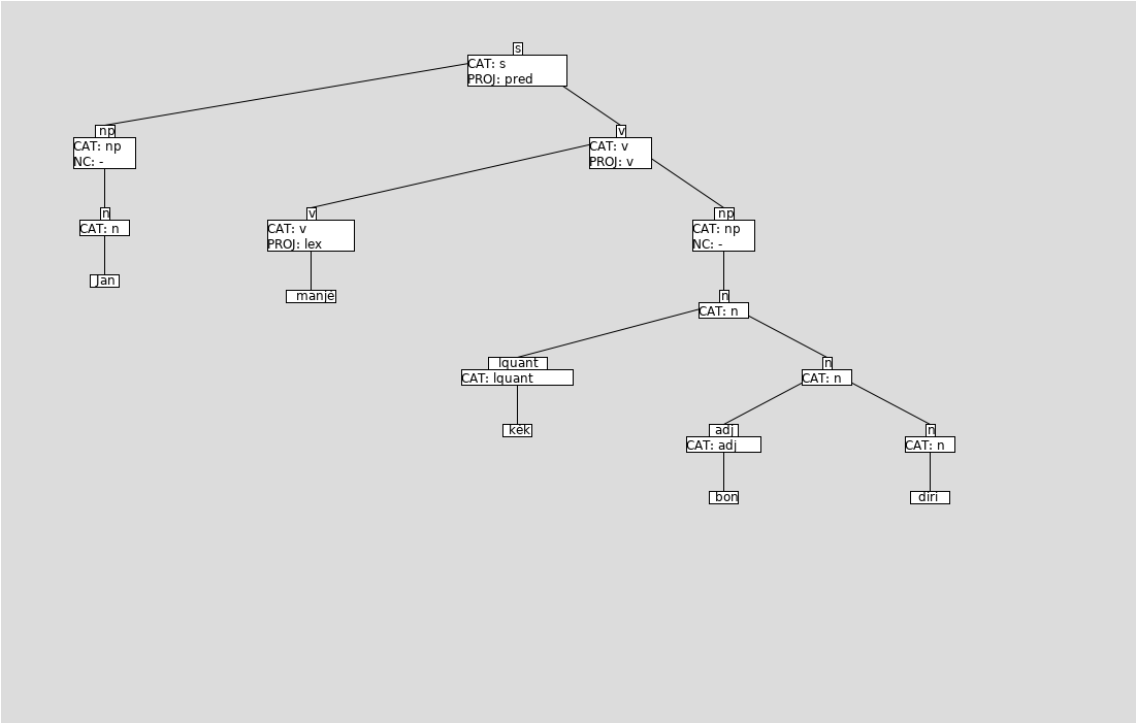


FIGURE 3.6 – Arbre dérivé de 'Jan manjé kèk bon diri'

3.4.6 Les séries verbales

Le titre de cette section fera probablement plaisir aux créolistes 'bickertoniens' mais c'est probablement le titre le plus trompeur de ce document... En effet, contrairement aux analyses faites dans l'APICS (https://apics-online.info/valuesets/50-86), je ne traite pas les constructions bénéfactives en (30) comme des séries verbales.

(30) Jan maké on lèt ban Mari
Jean écrire une lettre BA Marie
'Jean a écrit une lettre à Marie.'

En effet, même si étymologiquement *ba* dans cette position provient du verbe *ba*

'donner' (de l'A.Fr. *bailler* avec le même sens), en synchronie, il s'agit d'une préposition. Car :

— *ba* alterne librement dans cette construction avec *pou* 'pour' (*pou Mari*),

— *ba* n'accepte ici aucun TMA.

Je ne vois donc aucune raison de décrire ces structures comme des séries verbales.

Les constructions dites 'directionnelles' avec *alé* (31) sont également mentionnées dans l'APICS.

(31) Jan prann vwati a y (pou) alé Lapwent
Jean prendre voiture à lui ALLER Pointe-à-Pitre
' Jean a pris sa voiture pour aller à Pointe-à-Pitre'

Mes consultants préfèrent nettement utiliser 'pou alé Lapwent'. Je ne trouve donc pas convaincant le recours aux séries verbales pour ce créole. Je pense que Bickerton (1981), qui consacré beaucoup d'attention aux séries verbales en en faisant un trait caractéristique des créoles, en réaction à Muysken, Jansen, and Koopman (1978) a beaucoup (trop) influencé les descriptions des langues créoles sur ce point.

Je traite donc *ba* dans (30) comme une préposition, et cela ne pose aucun problème.

3.4.7 Conclusion

Les points exposés précédemment mettent en évidence le fait que la métagrammaire est à la fois un instrument 'pratique' permettant la création d'une ressource pour une langue de façon assez efficace, mais ils montrent également que la métagrammaire est un instrument au service de la théorie linguistique. En effet, l'analyse des TMA montre que le créole n'est pas dénué de morphologie verbale (v. Schang (2018) en particulier sur ce point). Bien au contraire, les TMA se trouvent à la frontière entre des éléments affixaux et des éléments périphrastiques. Les théories sur les créoles qui se fondent sur l'absence de morphologie

des créoles (à moins d'exclure de la catégorie les guadeloupéen et le santoméen) ne trouvent aucun appui dans ce travail. Je pense notamment ici à McWhorter (2011) qui oppose un tableau de conjugaison de l'espagnol et du palenquero (p. 33) afin de pouvoir affirmer que la morphologie des créoles est nécessairement plus simple que celle des autres langues.

La métagrammaire permet de modéliser également des fonctionnements assez complexes, comme le fonctionnement de la négation en guadeloupéen et les prédicats non verbaux. Cette modélisation rend compte du fait que la description de la négation en créole ne peut pas se limiter à un paramètre grossier (concordance négative stricte ou bien le fait que la négation soit toujours préverbale, cf. APICS, McWhorter (2011) ou encore Bakker et al. (2017)), et permet au passage d'illustrer le fait que *pa* est bien un élément négatif (contra l'analyse faire par Homer (Homer (2013)) pour le haïtien, selon qui il n'y a pas d'élément négatif 'ouvert').

La description des prédicats non verbaux (liée à celle des articles) illustre le fait que le guadeloupéen a un fonctionnement complexe dans le domain nominal (derrière la façade des noms nus qui seraient plus 'simples' que les SN du français par exemple) et que l'on ne saurait se réduire à affirmer que ce créole n'a pas de copule (cf. APCIS online et encore une fois McWhorter (2011)).

Enfin, cette métagrammaire montre que les séries verbales ne sont pas un concept nécessaire et utile pour la description du guadeloupéen. Le guadeloupéen, pas plus d'ailleurs que le santoméen [12], ne présente pas de séries verbales qui ne puissent être décrites comme un verbe suivi d'une préposition (même si celle-ci est étymologiquement issue d'un verbe). Cela rend ces créoles moins 'exotiques', mais la tâche du linguiste est de décrire les faits linguistiques, pas d'idéaliser les liens entre les langues créoles et un 'paradis perdu' africain.

12. Comme je l'avais déjà montré dans ma thèse à propos du santoméen (Schang (2000)).

Chapitre 4

Analyses sur corpus

Sommaire

4.1 Anaphores et coréférence 83
4.1.1 Retour en arrière 83
4.1.2 ANCOR 85
4.1.3 ANCOR-971 88
4.1.4 Contribution à l'analyse des chaînes de référence en créole 94
4.2 Temporalité 102
4.3 Intonation, prosodie 105
4.3.1 Tons, intonation et créoles 105
4.3.2 Question under discussion 108

4.1 Anaphores et coréférence

4.1.1 Retour en arrière

Pour un créoliste, le fait de travailler sur les anaphores en français n'est pas une évidence. Pourtant, mon parcours explique assez facilement pourquoi je consacre

une partie de mes recherches à ce sujet.

Lors de mes études à l'université Nancy 2, j'ai eu la chance de suivre les cours de Michel Charolles sur les anaphores (séminaire de DEA). Cela m'a permis d'être recruté au LORIA sur des projets liés à la coréférence et à la résolution des anaphores. Il s'agissait essentiellement de réaliser des annotations sur corpus en fonction d'un jeu d'annotations fixé par le projet. J'ai codé quelques relations anaphoriques dans *Le Père Goriot* (Balzac), mais surtout, au sein du projet Common-Ref, les relations anaphoriques autour des descriptions définies dans le Journal de la Communauté Européenne, en français et en portugais (à petite dose car c'était le travail des partenaires brésiliens essentiellement). Cela m'a permis de croiser Anne Reboul et de découvrir ses travaux en pragmatique, de connaître Renata Vieira (qui a travaillé avec M. Poesio) et de me familiariser avec les travaux de Massimo Poesio. Par la suite, j'ai contribué à annoter le corpus DéDé (Manuélian (2003)).

Le but principal était de financer mes études, bien entendu, mais j'ai surtout acquis des techniques (MMAX2, XML, etc.) et des méthodes de recherche qui m'étaient inconnues et que j'ai appris à maîtriser.

Ces travaux de linguistique de corpus m'ont conduit à penser qu'une approche quantitative des phénomènes linguistiques est complémentaire et nécessaire à une approche dite déductive (comme la grammaire générative par exemple). Pour énoncer les choses simplement : sans idées (sans intuitions), pas de schéma d'annotation pertinent, mais sans validation quantitative, pas d'analyse sérieuse.

Quelques années plus tard, cela m'a conduit à m'investir dans un WP du projet ANR Variling (porté par le LLL) qui a servi d'étude exploratoire pour d'autres projets, faute de données disponibles à temps.

J'ai porté un petit projet (3000 EUR, durée : 24 mois) associant le Laboratoire d'Informatique de Tours (LI) et le LLL. Ce projet a permis de tester notre capacité à travailler ensemble sur la tâche d'annotation du corpus ESLO (Enquêtes Socio-Linguistiques à Orléans). Nous avons testé plusieurs outils et guides d'annotations et, forts de cette expérience, nous avons décidé de candidater sur un appel à projet

régional.

4.1.2 ANCOR

J'ai été le porteur du projet ANCOR-Centre (APR-IA Région Centre sur 4 ans), largement guidé dans cette tâche par Jean-Yves Antoine (Laboratoire d'Informatique de Tours) pour toute la mise en oeuvre technique et la gestion de projet. Ce projet visait à constituer un corpus de taille importante annoté en anaphores et coréférences en français parlé. L'objectif a été atteint avec 488.000 mots (environ 30,5 heures de transcription d'enregistrements). Le contenu est le suivant :

CO2	35 000 mots - interviews sociolinguistiques
ESLO	417 000 mots - interviews sociolinguistiques
OTG	26 000 mots - dialogues interactifs, (office de tourisme)
Acccueil UBS	10 000 mots - dialogues interactifs, (standard université)

Une fiche descriptive précise de ce projet se trouve ici : https://www.ortolang.fr/market/corpora/ortolang-000903

Le corpus est constitué de différents sous-corpus (tableau ci-dessus). Nous avons utilisé les transcriptions comme base pour une annotation déportée réalisée avec le logiciel GLOZZ Widlöcher and Mathet (2012). Le schéma d'annotation est le fruit d'une collaboration entre le LLL et le LI et constitue donc un compromis entre des objectifs linguistiques et des objectifs applicatifs en traitement automatique. Ajoutons également le fait qu'un compromis a dû être fait avec le budget (toujours en dessous de ce qui est demandé dans le projet) et les compétences des annotateurs. Tout cela fait d'ANCOR un corpus certes imparfait (mais on pourrait se demander ce que serait un corpus parfait dans ce domaine) mais qui a le mérite d'exister et qui est un outil remarquable à bien des égards, comme je vais essayer de le montrer ci-dessous.

L'annotation est faite sur les transcriptions qui gardent les balises xml de Transcriber en leur sein. Les annotations sont faites dans un fichier .ac qui contient le repérage des entités et les relations entre ces entités.

Méthodologie d'annotation

L'annotation repose sur deux passages :

1. repérage des unités à annoter : l'annotateur repère les GN et leurs imbrications ; il décrit les traits morphosyntaxiques.
2. annotation des liens anaphoriques (relation Directe, Indirecte, Anaphore, Associative, Associative Pronominale).

Les traits repérés dans le corpus ANCOR

Le document qui présente en détail l'annotation de ce corpus est : http://tln.li.univ-tours.fr/Tln_Telechargement/Ancor-Centre.pdf

Problèmes

Les principaux problèmes que nous avons rencontré sont liés à des choix d'annotations (choix qui résultent de longues discussions au terme desquelles il a fallu trancher).

— annotation des relatives : les relatives ont été annotées par rapport à leur pronom et non incluses dans les GN. Comme indiqué dans le guide d'annotation :

> "Les pronoms relatifs ont un rôle très prévisible d'un point de vue syntaxique et sont le plus souvent connectés à leur antécédent. On pourrait donc choisir de s'affranchir de leur annotation. Dans certains cas, le pronom relatif peut être toutefois éloigné :
> — Il a acheté une maison près de Kercado qui est totalement à rénover
> — La maison bleue que j'aimais tant et dans laquelle j'ai passé les plus belles années de ma vie
> Dans ce cas, la caractérisation de la chaîne anaphorique n'est pas triviale sans l'annotation du pronom relatif. Par souci de cohérence, tous

> les pronoms relatifs seront donc considérés comme des entités [1]. Sur l'exemple 'J'ai une voiture qui est très rapide' on délimitera donc deux groupes nominaux :
> — une voiture
> — qui
> Et non pas un seul groupe nominal récursif : une voiture qui est très rapide."

Ce choix qui paraît prudent lors de la réalisation du guide d'annotation a été critiqué à plusieurs reprises (communications personnelles). Il reposait sur le fait qu'il est possible de rencontrer des éléments qui n'appartiennent pas au GN entre le pronom relatif et l'antécédent. Par exemple : "Jean a acheté une maison - Marie la pauvre en a pleuré toute la nuit - qui leur coûte les yeux de la tête". Finalement, (mais qui aurait pu le prédire), ces formes sont très rares. De plus, nous aurions pu utiliser pour celles-ci un outil de GLOZZ appelé 'schéma' comme nous l'avons fait pour les GN qui sont à cheval sur plusieurs tours de parole.

— *near-identity* : on trouve chez Recasens, Hovy, and Martí (2011) une proposition pour éviter à l'annotateur de trancher entre ce qui est coréférent et ce qui est *bridging* dans les cas peu commodes. Recasens, Hovy, and Martí (2011) affiche des taux d'accord inter-annotateur assez impressionnants et séduisants. Toutefois, lors d'une expérience d'annotation préliminaire nous n'avons pas constaté un tel succès, loin de là. Nous avons donc abandonné cette piste. Il semble que nous ayons eu raison de le faire car un projet polonais d'annotation des anaphores Ogrodniczuk et al. (2013) est arrivé au même constat que nous.

Conclusion

Ces quelques problèmes mis à part, il reste tout de même que ANCOR-Centre est le plus gros corpus annoté en chaînes de coréférence à l'heure actuelle pour

1. Il n'est pas évident cependant que 'qui' et 'que' soient des pronoms relatifs. Le tableau est probablement plus complexe, cf Abeillé and Godard (2007).

le français, en attendant le corpus de l'ANR DEMOCRAT[2]. Même si le corpus n'était pas annoté en chaînes au départ, il l'est désormais et permet des recherches à la fois sur la coréférence et sur les anaphores.

Je citerai ici Landragin (2017) pour conclure sur l'utilité de ce corpus :

> Pour la langue française, il n'existe à l'heure actuelle qu'un seul corpus permettant un apprentissage artificiel de qualité : le corpus ANCOR que nous avons déjà mentionné (Lefeuvre et al., 2014). C'est ce corpus que nous avons exploité pour réaliser une première expérimentation d'apprentissage artificiel, le système CROC – Coreference Resolution using Oral Corpus (Désoyer et al., 2014). Il ne s'agit cependant que d'une partie d'un système complet (ou « end-to-end »), c'est-à-dire d'un système capable de construire automatiquement les chaînes de coréférences rien qu'à partir du texte brut. En effet, pour fonctionner correctement, le système CROC doit partir d'un texte déjà annoté en formes de référence. Il ne correspond en quelque sorte qu'au dernier processus de la chaîne de traitement automatique, c'est -à-dire de la succession des opérations réalisées .

Récemment, le corpus ANCOR a servi de référence pour des études de TAL sur la coréférence : Brassier et al. (2018); Grobol (2019)

4.1.3 ANCOR-971

J'ai eu l'occasion d'avoir en cours une étudiante guadeloupéenne intéressée par le TAL qui a travaillé pendant un mois (contrat labo) à l'annotation en chaînes de référence sur le seul corpus de guadeloupéen disponible : Glaude (2013). Ceci m'a permis d'obtenir un corpus d'environ 3h d'enregistrements transcrits annotés en coréférences.

Je vais détailler un peu le contenu de ce (micro-) projet qui, bien que proche du projet ANCOR-Centre, possède quelques particularités liées au créole.

2. Le corpus est désormais disponible : https://www.ortolang.fr/market/corpora/democrat/v1

Expressions référentielles et créoles

La question qui occupe principalement les créolistes, comme en témoigne Baptista and Guéron (2007) est celle des SN sans déterminants (Bare NPs). Les travaux tournent autour des préoccupations exprimées par la grammaire générative essentiellement. En l'absence de déterminant, quelle est la forme du DP ? Existe-t-il des contraintes syntaxiques sur l'interprétation des bare NPs ? Quelles projections doit-on postuler au sein du DP ?

On peut cependant poser la question des Bare NPs autrement, comme le fait d'ailleurs Déprez (2005), qui n'hésite pas à prendre ses exemples en corpus ('Le Petit Prince' en l'occurrence).

Je rapproche bien entendu ce travail sur le créole des travaux sur le corpus ANCOR et j'adopte une approche 'TAL' de ce problème : soit une mention dans le discours, cette mention est-elle celle d'une nouvelle entité du discours ou bien doit-elle être interprétée comme pointant vers une entité déjà mentionnée dans le discours ? Et si c'est le cas, laquelle ?

Comme nous le notions dans Antoine, Lefeuvre, and Schang (2016) :

> La résolution des anaphores est un sujet ancien et central dans le traitement automatique du discours (v. Mitkov 2010 notamment pour une présentation et un état de l'art). Depuis Karttunen (1976) on envisage essentiellement le problème de la manière suivante :
>
> « Consider a device designed to read a text in some natural language, interpret it, and store the content in some manner, say, for the purpose of being able to answer questions about it. To accomplish this task, the machine will have to fulfil at least the following basic requirement. It has to be able to build a file that consists of records of all the individuals, that is, events, objects, etc., mentioned in the text and, for each individual, record whatever is said about it. »
>
> Mais déterminer comment il est fait mention des entités (individuals) dans le discours n'est pas une tâche simple. A la différence des langages mathématiques (par exemple), les langues naturelles n'assignent

> pas un identifiant unique aux objets auxquels il est fait mention par les locuteurs. On appelle chaîne de références la suite des expressions d'un texte qui ont la même identité référentielle (Schnedecker & Landragin 2014). La tâche qui consiste à identifier ces chaînes est difficile à automatiser et demande des corpus annotés en coréférence afin de dégager des heuristiques de traitement ou pour l'entraînement des approches par apprentissage automatique.

La question ainsi posée est donc très différente de celle qui est habituellement posée à la suite de Russell (1905) que l'on peut énoncer ainsi : "quelles sont les conditions de vérité d'un énoncé en fonction des déterminants d'un NP ?"

Contrairement aux méthodes que je vais appeler 'classiques', il s'agit ici 1) de prédire une forme de reprise d'une mention, 2) de relier ou non une mention à un antécédent.

Il s'agit donc de bâtir un modèle général permettant de prédire des faits. Dans ce cadre, les *bare NPs* (BNP) sont décrits dans leurs usages sans chercher à relier ces usages à des projections syntaxiques particulières censées expliquer les différents usages.

La principale critique que l'on peut faire aux modèles 'classiques' (pour être explicite, disons les approches semblables à Zribi-Hertz (1996)) est que ce sont des approches locales de la coréférence. La grammaire générative ne présentant aucun modèle dépassant les limites de la phrase, le lien avec le niveau discursif est peu, ou pas, explicité. L'interface avec le discours est pourtant souvent évoquée comme une pièce cruciale de l'analyse (v. Aboh (2015) par exemple) mais il n'existe pas, à ma connaissance, de modèle explicite de cette interface et du lien avec la pragmatique du discours.

Or, les études sur corpus montrent que la résolution des phénomènes anaphoriques ou/et de coréférence ne se limitent pas à la phrase ou à la succession de deux phrases (comme dans la théorie du Centrage, v. Walker, Joshi, and Prince (1998)). C'est également ce que je vais montrer ci-après.

Description de l'annotation du corpus

Le corpus transcrit est mis sous la forme de textes en entrée de l'outil d'annotation GLOZZ (Widlöcher and Mathet (2012)). La DTD qui représente le schéma d'annotation est fournie aux annotateurs ainsi qu'un guide d'annotation.

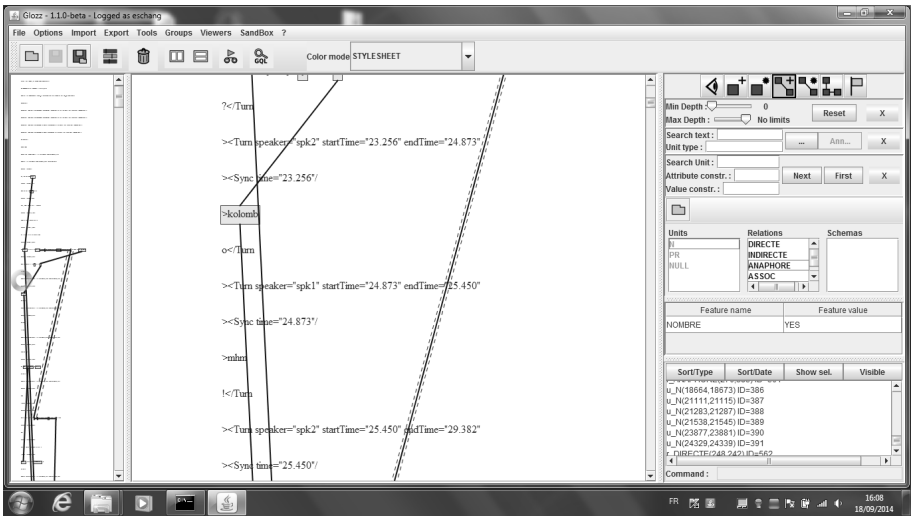


FIGURE 4.1 – L'interface utilisateur de GLOZZ

Dans une première étape, l'annotateur (ou les annotateurs) identifient les mentions (les GN) et donnent leur type :

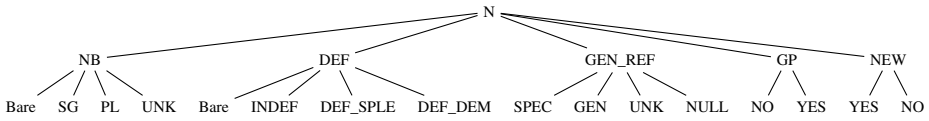
— catégorie : (N ou Pr),
— fonction syntaxique,
— définitude (bare, indéfini, etc.),
— spécificité (ou généricité) [3],
— inclusion dans un groupe prépositionnel (utile pour repérer les N de N)
— la nouveauté en discours (discourse new / discourse old).

L'arbre xml ci-dessous illustre la façon d'ordonner les traits :

3. Ce trait s'avèrera inutile car l'annotation est assez peu fiable, les annotateurs étant très indécis sur des exemples non stéréotypés.



FIGURE 4.2 – Visualisation des chaînes avec GLOZZ



Dans un second temps, l'annotateur lie les mentions entre elles (si nécessaire) et donne le type de la relation :

— directe : la reprise et l'antécédent ont la même tête nominale.

(1) i pwéparé on *kolombo* … *kolombo la* té bon
3SG préparer un colombo … colombo PST bon
'il a préparé un colombo … le colombo était bon'

— indirecte (ou infidèle) : relation de coréférence mais la tête nominale est différente.

(2) i pwéparé on *kolombo* … *manjé -la* té bon
3SG préparer un colombo … DEF PST bon
'il a préparé un colombo … ce plat était bon'

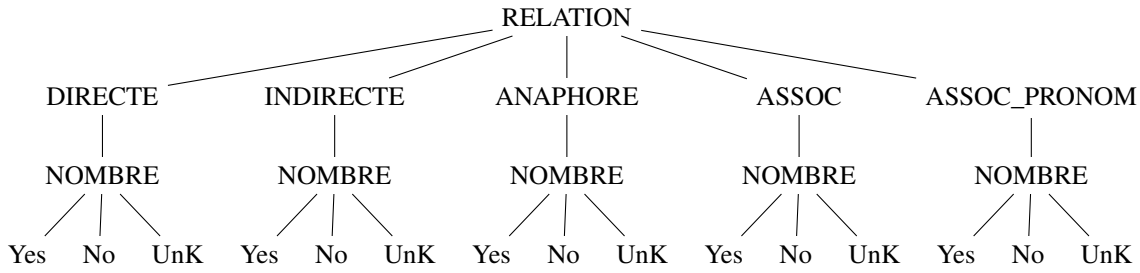
— anaphore : reprise par un pronom.

(3) i pwéparé on *kolombo* ... *i* té bon
3SG préparer un colombo ... 3SG PST bon
'il a préparé un colombo ... il était bon'

— associative et associative pronominale : il s'agit des cas d'anaphore sans coréférence (bridging anaphora en anglais).

Un trait Nombre a été conservé du schéma réalisé pour ANCOR. Il s'agit de repérer les relations dans lesquelles le nombre grammatical est différent (la foule... ils). Finalement, ce trait n'est pas fiable en raison du nombre élévé de SN sans déterminants pour lesquels l'attribution du nombre est très discutable.

Le schéma est donc le suivant :



La structuration des traits est un élément clé de l'annotation. J'ai bénéficié ici de l'expérience d'ANCOR mais il faut avouer que tout cela reste encore perfectible.

A la différence d'ANCOR, qui demandait à l'annotateur de pointer la relation sur la première mention d'une entité, il était demandé ici de pointer vers l'antécédent le plus proche. Ce choix permet de travailler directement sur des chaînes de référence (v. Schnedecker and Landragin (2014)). C'était l'objectif principal de ce mini-projet.

Des exemples pris dans ce corpus illustrent la documentation de la norme ISO-

RAF (reference annotation framework) qui est en cours de rédaction [4].

Outils pour l'analyse

GLOZZ dispose d'un outil d'analyse intégré, GLOZZ-QL, qui permet de visualiser les données annotées. Cependant, cet outil ne fonctionne que fichier par fichier et non pas sur un corpus entier (au moins dans la version utilisée alors).

Pour pouvoir analyser rapidement le corpus, j'ai utilisé ANCOR-QI (Lefeuvre, Antoine, and Schang (2014)) qui a été développé par Anaïs Lefeuvre pour le corpus ANCOR. Cet outil permet des requêtes intéressantes et dispose d'un concordancier. Cependant, il n'est pas possible de visualiser les données avec ANCOR-QI.

Sur les conseils d'Olivier Bonami, j'ai appris le langage Python et surtout l'utilisation des bibliothèques Pandas, Matplotlib, Altair, etc. qui me permettent de lancer des scripts sur le corpus en entier et de visualiser les données efficacement, sans dépendre d'outils particuliers.

4.1.4 Contribution à l'analyse des chaînes de référence en créole

Cette partie se comprend par rapport au travail effectué dans ANCOR. Les questions posées sont les suivantes :

— étant donné que le créole ne possède pas de genre grammatical, est-ce que les chaînes de référence sont différentes de celles du français ? Est-ce que les algorithmes (ou le jeu de traits) pour la résolution de la coréférence et des anaphores peuvent reposer sur les mêmes bases ? La question est d'autant plus intéressante que le créole et le français partagent un grand nombre de mots (par delà les dissemblances graphiques, *manjé* et *manger*, *pwason* et *poisson* ont le même sens et la même prononciation).

4. ISO/DIS 24617-9 Language resource management – Semantic annotation framework (SemAF) – Part 9 : Reference annotation framework (RAF), Project leaders : Laurent Romary and Kiyong Lee

— sachant qu'il existe des SN créoles dans lesquels il n'y a aucun déterminant (SNSD), est-ce que ce trait a un impact sur les chaînes de référence ?

Je livre ci-dessous quelques résultats issus de Schang, Antoine, and Lefeuvre-Halftermeyer (2017) et d'un article soumis (mêmes auteurs) à la Revue de Sémantique et Pragmatique (RSP).

Les mentions

Le corpus ANCOR-971 contient 2731 entités et 1225 relations, ce qui représente une base de travail considérablement plus large par rapport aux études antérieures sur le créole. A titre d'exemple, Gadelii (2007) a travaillé sur un corpus contenant cent exemples.

On compte 1905 N et 814 PR. Ces chiffres sont étonnants si l'on compare avec le corpus ANCOR-Centre (français) où les N constituent la moitié des entités.

TABLE 4.1 – Proportion de SN vs Pronoms dans les deux langues

	N	Pr
Français	51.2%	48.8%
Créole	70%	30%

On remarque que les SN qui n'ont pas de déterminant (ni défini, ni démonstratif, ni pluriel, ni indéfini) représentent la majorité des mentions.

Le décompte des SN par type de définitude montre que les SNSD sont de loin les plus nombreux :

— SNSD (bare NPs) : 1078
— Indéfinis (INDEF) : 401
— Définis (DEF SPLE) : 398
— Démonstratifs (DEF DEM) : 28

Dans ce contexte, il est crucial de pouvoir trouver une heuristique de résolution des coréférence qui prenne efficacement en compte les SN dans lesquels ne figurent pas de déterminant, et contrairement au français où les occurrences de SNSD référentiels sont réduites (cf. Bouchard (2003) notamment).

La première question que nous nous sommes posée concerne la nouveauté dans le discours. Les SNSD sont-ils plus utilisés pour introduire de nouvelles entités du discours (NEW) ou bien comme reprise (OLD) ?

On remarque en figure 4.3 et figure 4.4 que les SNSD et les définis ne s'opposent pas sur le trait (NEW/OLD). Ils se distribuent en proportion similaire sur ce critère, en revanche, bien entendu, le nombre d'occurrences n'est pas le même, les SNSD étant de loin plus nombreux. Sans surprise, les indéfinis sont utilisés majoritairement pour introduire des nouvelles entités (NEW), comme en français.

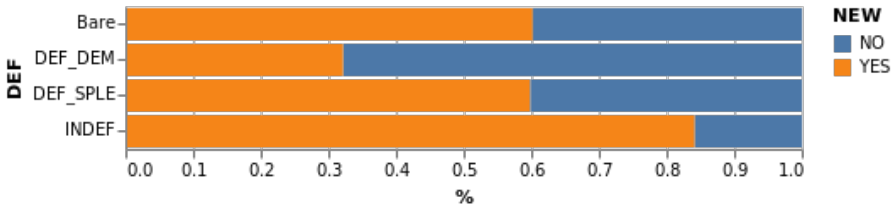


FIGURE 4.3 – Distribution (%) de la nouveauté suivant les différents type de définitude des mentions nominales.

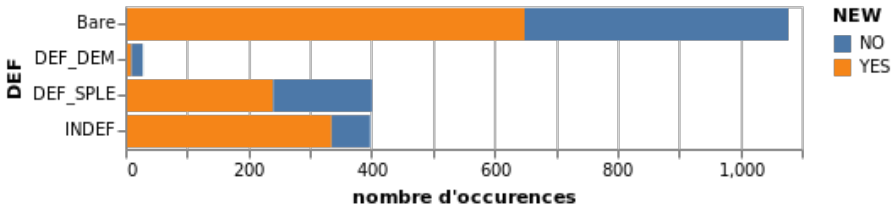


FIGURE 4.4 – Distribution de la nouveauté suivant les différents type de définitude des mentions nominales

On peut se poser également la question des 'singletons' (mentions non reprises) et qui constituent un challenge important pour les solveurs de coréférence (v. De Marneffe, Recasens, and Potts (2015); Recasens, de Marneffe, and Potts (2013) notamment). En effet, peut-on trouver une différence dans la proportion des singletons entre les SNSD et les définis ?

La réponse est non, ou en tout cas, les données de ce corpus ne permettent

pas d'appuyer cette idée, car on trouve 66% de SNSD dans les singletons (donc jamais repris dans le discours) contre 63% pour les définis.

Les autres critères, comme la fonction syntaxique par exemple, n'apportent pas non plus de différence flagrante entre les catégories.

Les relations

Les relations de coréférence annotées se distribuent de la manière suivante :

— ANAPHORE : 604

— DIRECTE : 402

— INDIRECTE : 102

La figure 4.5 illustre la répartition en pourcentage de l'ensemble des relations (y compris les associatives). Il est intéressant de remarquer que les reprises 'directes' (reprise par une même tête nominale, aussi appelée 'anaphore fidèle') représentent une part importante des relations, et la répartition des relations est assez proche de celle trouvée dans le corpus ANCOR-Centre pour le français (41,6% d'anaphores pronominales et 39,9% de relations directes, v. Muzerelle et al. (2012); Lefeuvre, Antoine, and Schang (2014)).

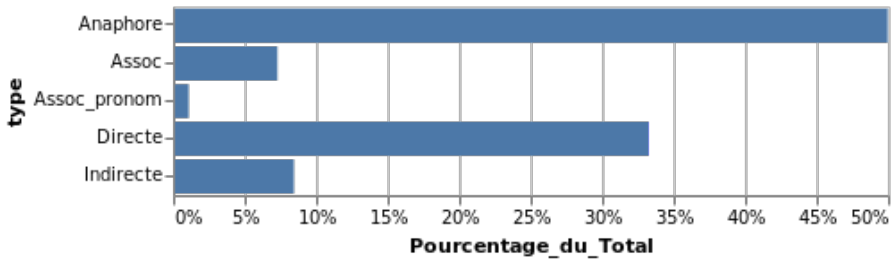


FIGURE 4.5 – Proportion des différentes relations

La principale différence entre le français oral (ANCOR-Centre) et ce corpus de créole tient dans la structure des chaînes de coréférence. En effet, en français, la chaîne prototypique (la chaîne la plus probable distributionnellement) est de type N-N-N-P-P [5] (v. Antoine, Lefeuvre, and Schang (2016)). Deux types de chaînes

5. On note N à la place de SN pour mettre en évidence le fait que la tête du groupe est un nom.

apparaissent prioritairement en créole avec une ancre nominale : N-P-P-P et N-N-N-N. La figure 4.6 illustre par un graphe la structure des chaînes coréférentielles ancrées par un N (Na).

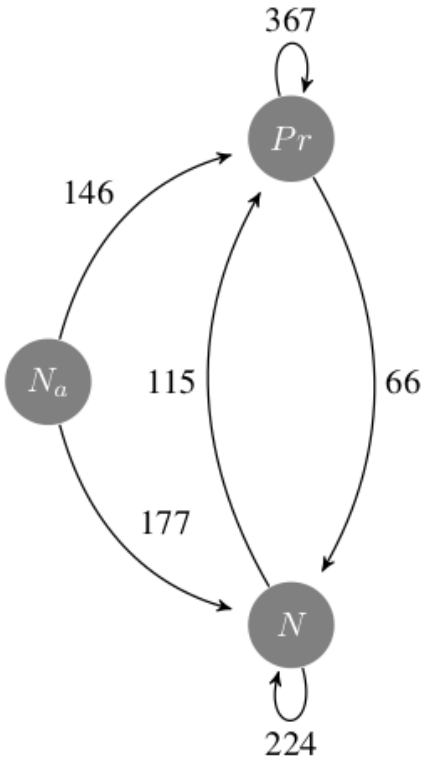


FIGURE 4.6 – Chaînes coréférentielles ancrées par un N (Na) par type et en nombre de relations.

A ce stade, il est difficile d'interprêter cette différence. Il n'est pas exclu que ce soit une particularité du corpus recueilli.

En revanche, cette étude a mis en évidence un point qui me paraît essentiel : la forme de l'élément en reprise (pronom, SNSD, défini) est sensible à la distance avec l'antécédent, mesurée en nombre de mentions entre les deux éléments. Dans l'exemple ci-dessous, l'antécédent et la reprise sont distants de 3 car 3 mentions

qui n'appartiennent pas à leur chaîne de coréférence les séparent dans le discours. On peut considérer que plus il y a de mentions hors chaîne entre deux maillons d'une chaîne, plus l'accessiblité de l'antécédent est dégradée.

```
antécédent... N... Pr... N... reprise
```

Même si cette distance est évoquée dans la littérature sur la coréférence, notamment du point de vue de la saillance et l'accessibilité du référent (voir notamment Gundel, Hedberg, and Zacharski (2001); Gundel (2003, 2010); Gundel, Hegarty, and Borthen (2003)), ce critère est absent de la plupart des analyses, que ce soit sur le français ou le créole.

Or, il semblerait que ce critère soit pertinent pour le choix de l'expression en reprise, et notamment pour les SNSD.

En effet, la figure 4.7 présente la distance entre un pronom et son antécédent ancre nominale (en nombre de mentions intermédiaires) et par type de définitude. On remarque que les pronoms reprennent localement un antécédent ancre. La plupart des pronoms ont un antécédent situé à deux mentions ou moins de distance. Les quelques points éloignés (distance >10) sont des cas de pronoms dont l'antécédent est le thème général du discours. Rien n'indique donc que ce soit véritablement ces antécédents discursifs qu'ils reprennent.

Si le pronom n'est pas en second, mais plus loin dans la chaîne, on observe en figure 4.8 que les antécédents sont majoritairement des définis.

Lorsqu'un pronom reprend un pronom (maillon Pr-Pr), la distance médiane est de 1 entité et 75% des maillons Pr-Pr ont un maximum de deux entités d'écart. On peut donc en déduire que la résolution des pronoms est bien une question locale. Les pronoms qui ont un antécédent éloigné sont peu nombreux et leur attachement référentiel est à trouver ailleurs que dans le texte.

Schang, Antoine, and Lefeuvre-Halftermeyer (2017) montre que les SNSD ont un antécédent moins proche, en moyenne, que les pronoms. En effet, la distance moyenne entre un pronom et son antécédent (en nombre d'entités présentes entre les deux dans le discours) est de 3.07 (médiane = 1). Pour les SNSD, la distance

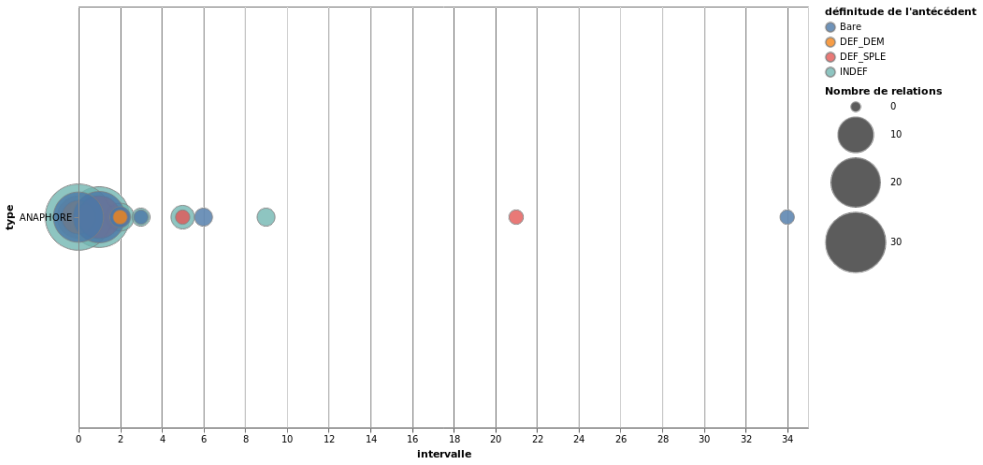


FIGURE 4.7 – Distance entre un pronom en seconde mention et son antécédent (par définitude).

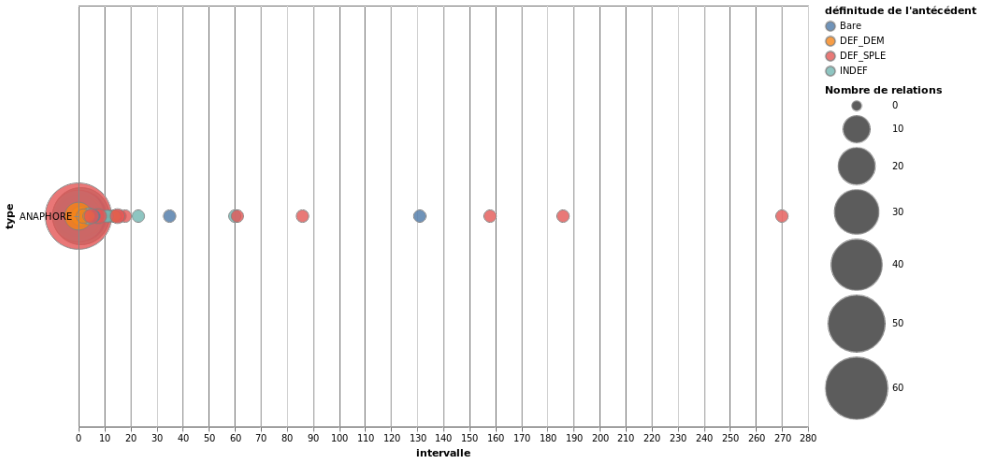


FIGURE 4.8 – Distance entre un pronom et son antécédent non-ancre (par définitude).

moyenne est de 8,83 (médiane =2). Les définis (principalement spécifiques en créole) se trouvent encore plus distants : distance moyenne = 14, médiane = 4.

Encore une fois, il faut être prudent car il faudrait plus de données pour pouvoir conclure de manière certaine, cependant ces analyses permettent de faire des hypothèses sur l'usage des SN, si nombreux en créole par rapport au français, et sur l'absence de genre.

En effet, on peut faire l'hypothèse que le créole a recours aux SNSD comme stratégie permettant de pallier l'absence de genre. En effet, la plupart des analyses (v. Blackburn (2005) par exemple) mettent en avant le genre comme étant un trait essentiel dans la résolution des anaphores. En l'absence de genre, la stratégie choisie par le créole est donc de faire une reprise par un SNSD ayant la même tête lexicale que l'antécédent, comme le montre le taux élevé de reprises directes. Les pronoms sont alors principalement utilisés 'localement', c'est-à-dire lorsqu'il existe peu de mentions hors chaîne qui pourraient mener à de mauvaises interprétations.

Dans Schang, Antoine, and Lefeuvre-Halftermeyer (2017) nous avons proposé l'analyse suivante :

> "On peut faire l'hypothèse raisonnable que, en l'absence de genre grammatical, les pronoms sont utilisés pour reprendre une entité du discours immédiatement disponible et, dès lors qu'une ambiguïté est possible, la reprise par un SNSD est préférée. On remarque par ailleurs que les SN définis trouvent leur antécédent à une distance (en nombre de mentions) moyenne de 14 (médiane = 4). On le voit donc, ils reprennent en moyenne des mentions situées plus loin que les SNSD."

Ceci mériterait d'être vérifié sur des corpus plus grands. Il s'agit en tout cas d'une piste de recherche intéressante. En effet, elle met en difficulté toutes les théories sur la coréférence qui ne prennent pas en compte la 'profondeur' du discours. Je pense notamment aux théories qui réduisent le contexte à la phrase précédente (comme la théorie du Centrage notamment Walker, Joshi, and Prince (1998)) ou bien celles qui présentent un stock non ordonné de référents dispo-

nibles dans lequel il faut piocher un antécédent, comme Reboul and Moeschler (1998) par exemple [6].

4.2 Temporalité

Après avoir étudié les expressions anaphoriques et les relations de coréférence, il m'a semblé intéressant de chercher à développer des annotations complémentaires sur les expressions temporelles. En effet, on compte dans ANCOR de nombreux SN qui renvoient à des référents temporels. Le schéma d'annotation d'ANCOR n'étant pas conçu pour prendre en compte spécifiquement les événements, nous avons convenu de monter un projet appelé TEMPORAL (suivi par la suite d'une extension appelée TEMPORAL@ODIL qui constitue un work package d'un autre projet APR-IA). Jean-Yves Antoine en a assuré la coordination et j'ai assuré la coordination de la tâche 2 (annotation de corpus).

Ce projet a associé des linguistes du LLL et des informaticiens du LI de Tours et du LIFO à Orléans. Les travaux ont permis de réaliser l'annotation du corpus ANCOR (une sous-partie de 10 000 mots) en arbres afin de pouvoir annoter les noeuds de ces arbres. Nous avons travaillé à un travail critique sur la norme ISO-TimeML qui ne nous paraît pas correspondre aux attentes ni des linguistes, ni des informaticiens travaillant sur des langues autres que l'anglais Lefeuvre et al. (2014); Antoine et al. (2017); Lefeuvre-Halftermeyer et al. (2016) [7].

Un outil d'annotation d'arbres a été conçu pour ce projet : CONTEMPLATA (https://github.com/kawu/contemplata) par Jakub Waszczuk lors de son travail d'ingénieur dans ce projet. Dans ce cadre, le Stanford parser a été entraîné sur le French Treebank (Abeillé (2011)) [8]) et nous a servi à produire les arbres correspondant au corpus ANCOR. S'agissant d'un corpus oral, on s'attendait à

6. Je ne mentionne ici que des théories avec un fort ancrage sémantico-pragmatique, mais il va de soi que certaines approches 'génératives' de l'anaphore sont totalement incompatibles avec ces analyses, cf. Zribi-Hertz (1996) par exemple

7. Voir également Bittar et al. (2011) pour une approche différente de la nôtre.

8. http://ftb.linguist.univ-paris-diderot.fr/

des résultats perfectibles. C'est pourquoi l'outil permet de corriger facilement les arbres produits et de relancer le parser en intégrant les modifications (comme montré dans la figure 4.9).

Je ne vais pas détailler plus avant ce projet (qui s'achèvera début 2019), mais je vais expliquer en quoi cela m'intéresse comme créoliste.

Comme l'explique fort justement Winford (2012), la créolistique s'intéresse fortement à la question du marquage grammatical du temps depuis que Bickerton (1981) en a fait une question centrale de la créolisation.

Il est évident, comme le montre l'exemple suivant, que les langues créoles encodent grammaticalement le temps et l'aspect d'une manière différente de leur langue lexificatrice.

(4) a. Jan ka manjé kolonbo
Jean IPFV manger colombo
'Jean mange du colombo'

b. Jan ké manjé kolonbo
Jean PROSP manger colombo
'Jean mangera du colombo'

De nombreux travaux ont décrit les marqueurs TMA (Temps, Mode et Aspect) des langues créoles et j'ai également contribué à la discussion sur ces marqueurs (Schang (2002); Schang et al. (2012a); Schang (2013)). Toutefois, les conditions d'emploi des marqueurs dans les relations temporelles restent assez mal définies (bien que Pfänder (2000) constitue une avancée significative). Nous ne disposons pas de corpus annotés en relations temporelles pour les créoles, ce qui pourtant serait très intéressant, notamment dans une optique de comparaison avec le français.

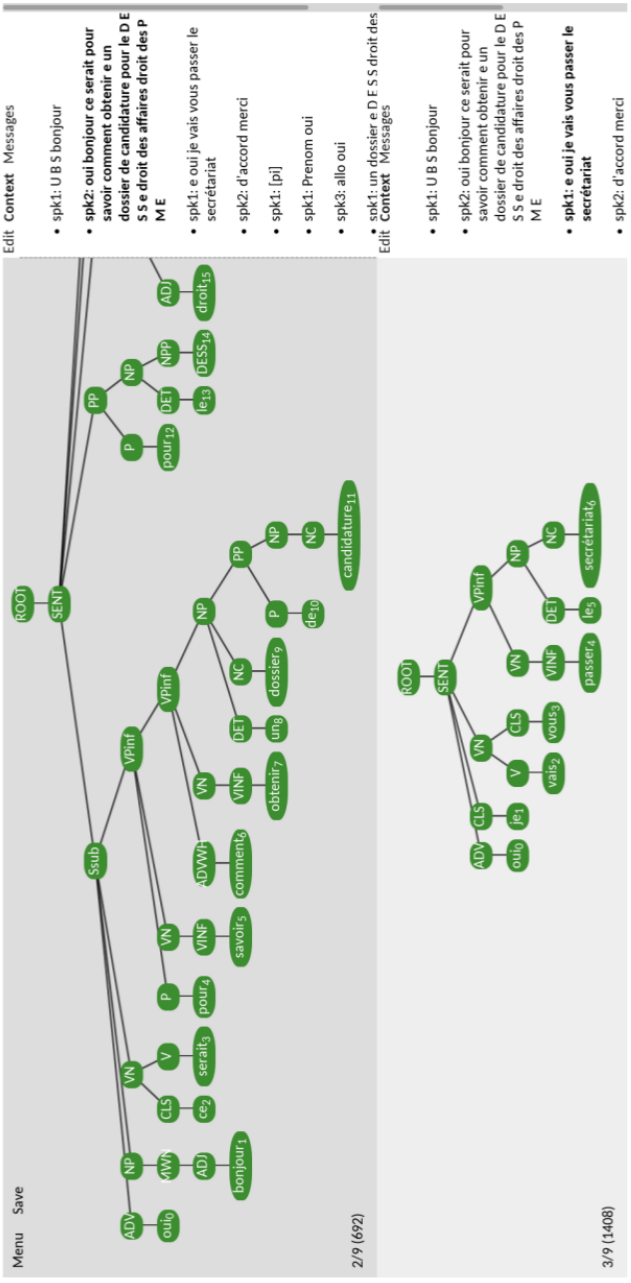


FIGURE 4.9 – L'interface web de CONTEMPLATA

4.3 Intonation, prosodie

4.3.1 Tons, intonation et créoles

La question de la prosodie, de l'intonation, de l'accent et des tons est une question à la fois souvent évoquée dans les études créoles et en même temps assez mal traitée[9]. On dispose finalement d'assez peu de matière sur le sujet, faute de corpus disponibles probablement.

Si cette question est importante, c'est parce qu'elle est cruciale pour plusieurs théories sur l'émergence des créoles :

— si les créoles sont des langues tonales, ils se rapprochent des langues 'africaines' qui composent le substrat (le fongbe par exemple),

— si elles n'ont pas de tons, elles sont simples (cf. McWhorter (2001)) et par conséquent, certains tiennent à démontrer qu'elles n'ont pas de tons.

Cette vision est bien entendu simpliste. En effet, il resterait à prouver que les langues accentuelles sont plus 'simples' que les langues tonales, ce qui, si on connaît les difficultés à décrire (modéliser) l'accent en français et en anglais, est loin d'être évident.

Je n'ai pas travaillé directement sur la question des tons dans les créoles du Golfe de Guinée après ma thèse. A l'époque, j'avais défendu une vision différente de celle de Philippe Maurer (Maurer (2008, 1995)) et de Tjerk Hagemeijer (communications à l'ACBLPE) qui défendaient (ou défendent encore) l'existence de tons en angolar et en forro. Si T. Hagemeijer ne défend plus ces positions, elles restent toutefois fortement soutenues par Ph. Maurer (Michaelis et al. (2013)) :

> Ferraz (1979) claims that Santome lacks phonologically signficant tone, although it is used as a stylistic or emphatic device. Maurer (2008), however, argues that Santome has a simple tone system with a two-way contrast. The examples are taken from Maurer (2008).
>
> Confidence : Very certain

9. Je pense que cette question est d'ailleurs encore en débat pour de nombreuses langues.

Les exemples donnés en illustration sont supposés démontrer que le forro est une langue tonale. Je continue cependant à penser, comme Ferraz (1978)), que pour qu'il y ait des tons, il faut qu'il y ait de véritables paires phonologiques, ce qui n'est pas montré dans les exemples donnés.

En effet, les paires lexicales données en illustration ne sont pas reconnues par les locuteurs. Par ailleurs, la paire censée illustrer une opposition entre un irréaliste *ka* et un *ka* habituel (ma traduction) ci-dessous est plus que douteuse dans la mesure où l'irréaliste est marqué par *xi* et non par *ka*.

(5) Xi tudu mwala myole ká sêbê kwa se
si chaque femme aujourd'hui IRR savoir chose DEM [...]
'Si toutes les femmes d'aujourd'hui savaient cette chose [...].'

L'analyse défendue repose sur une vision tranchée des phénomènes prosodiques qui repose sur le lien bi-univoque entre ton/accent et syllabe. Les phénomènes sont ainsi rigoureusement circonscrits au seul domaine syllabique. Cela ne va pas sans poser de gros problèmes de description, comme nous allons le voir.

On notera que l'analyse de Ph. Maurer est critiquée également par dos Santos Agostinho (2016) pour le lung'Ie. Mais ici également, à part montrer que l'analyse tonale est peu convaincante, il n'y a pas vraiment d'analyse concurrente.

Pour poser le problème de façon simple (de manière à peine caricaturale), l'enjeu est de dépasser une description qui ferait du français une langue tonale car on aurait trouvé dans des enregistrements une différence de hauteur sur *venait* dans les exemples suivants :

(6) a. Jean venait tous les jours. (Habituel)
b. Si Jean venait, Marie serait contente. (Irréaliste)

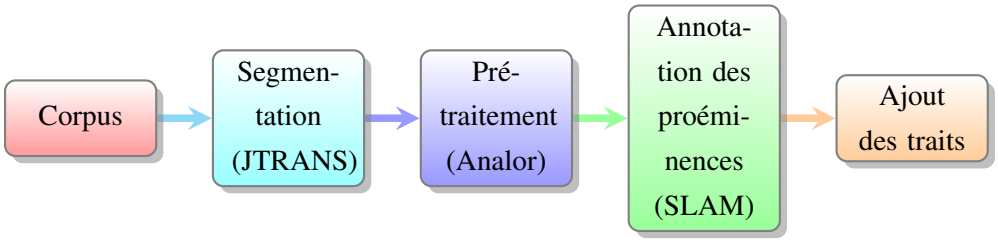
Cette différence de hauteur que l'on constatera probablement en français ne permet pas (sauf confusion théorique) de déduire que le français est une langue tonale, avec des tons grammaticaux permettant de distinguer l'Habituel de l'Irréaliste. C'est pourtant cette méthode qui est appliquée aux créoles du Golfe de Guinée…

On remarquera d'ailleurs que le dictionnaire de Jean-Louis Rougé (Rougé (2004)) ne fait aucunement référence à des tons dans ces créoles. Ceci amène tout de même à s'interroger : s'il est possible de faire un dictionnaire contenant plusieurs milliers d'entrées sur ces langues sans décrire les tons, n'est-ce pas parce que ces tons n'existent pas (phonologiquement) ?

La question qui se pose donc est celle d'établir une méthodologie qui permette de dépasser les quelques exemples construits pour aller vers une annotation complète de corpus. Et ceci dans une approche guidée par le corpus, sans a priori théorique sur le lien entre la syllabe et l'intonation.

Pour ce faire, on ne peut pas s'appuyer sur une annotation manuelle des tons et accents, trop hasardeuse comme l'ont montré Morel et al. (2006). Pour contourner ce problème, je travaille à l'élaboration d'une chaine de traitement, avec l'aide de Flora Badin (IE du LLL) qui parte de textes transcrits pour réaliser une annotation des proéminences avec le logiciel SLAM Obin et al. (2014).

La procédure est la suivante [10] :



Cela permet de lancer des scripts python qui détectent les corrélations entre certains contours prosodiques et les unités à annoter (les expressions référentielles dans ANCOR et ANCOR-971 par exemple).

Cette procédure est encore en test car elle demande des ajustements fins sur les frontières de groupes, sur l'alignement et sur les paramètres de l'analyse du signal. Elle demande également de travailler locuteur par locuteur afin d'établir un diapason pour chaque locuteur sur lequel les analyses vont être produites (calcul d'un différentiel par rapport au diapason).

10. Cette procédure a été présentée aux journées FLORAL (2017) sous le titre "Automatiser l'analyse prosodique des corpus oraux" (Badin Flora, Schang Emmanuel, Nemo François & Leroux Camille).

Un des points à améliorer est la qualité du signal sonore car les enregistrements ont été faits dans des conditions imparfaites, ce qui demande un travail de prétraitement des courbes de F0.

SLAM permet de modéliser un contour prosodique à l'aide d'un vocabulaire réduit :

— H,h,m,l,L pour les hauteurs,

— 1, 2, 3 pour la place d'une proéminence dans le segment concerné (exprimée en tiers de segment.

Un segment annoté *mmh3* aura donc le contour suivant :

— départ au niveau médian,

— fin au niveau médian,

— proéminence de niveau haut (h) dans le 3eme tiers du segment.

La figure 4.10 présente un exemple de traitement sur le GP 'au garage' et qui est de contour *mm*.

Même si ce travail n'est pas encore complètement finalisé, il se révèle prometteur car il permet de dépasser l'annotation manuelle et de travailler sur des données massives. C'est d'ailleurs l'un des intérêts d'inciter à la collecte et à la mise à disposition des corpus oraux de pouvoir adopter cette méthodologie.

4.3.2 Question under discussion

Dans un atelier de la DGfS 2017 sur la prosodie, j'ai présenté avec F. Nemo et F. Krimou une communication intitulée 'Uttered Sentences, Prosody and Word Order' qui montrait qu'une approche naïve liant la syntaxe, la prosodie et un type illocutoire donné ne permet pas de rendre compte des faits constatés en français et en anglais.

Nous avons montré que la portée de la prosodie n'est pas l'énoncé (encore moins la phrase) mais la contribution (niveau discursif) et nous concluons ainsi :

"What is the semantic function of prosody ?

— Not only about sentence type

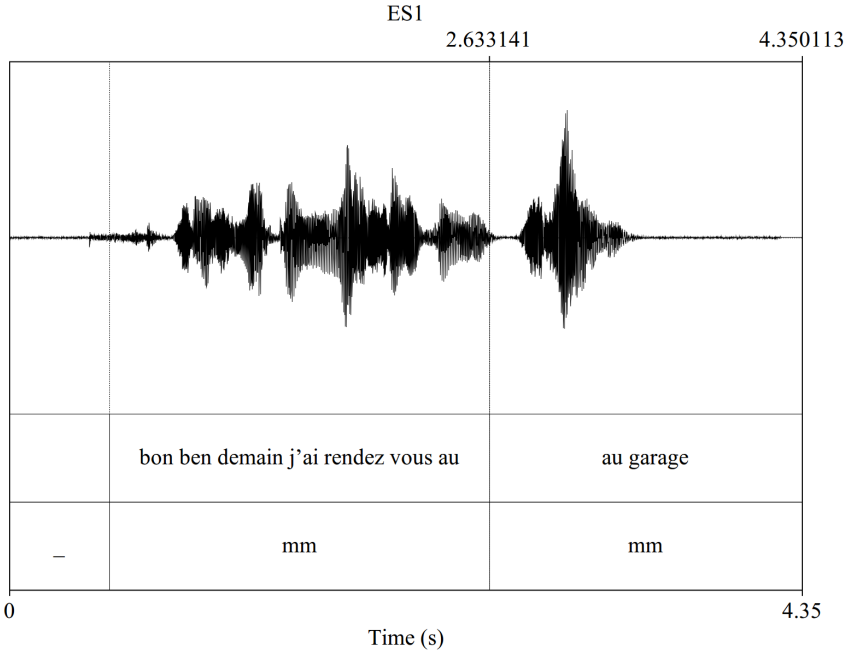


FIGURE 4.10 – Un TextGrid de PRAAT présentant une modélisation du contour prosodique (mm : médian-médian) réalisée avec SLAM.

— No complete redundancy with word order
— Not only illocutionary value

Prosody is providing information about the issue at stake which can be the Question Under Discussion."

Ainsi, on peut voir sur la figure 4.11 que le même syntagme [3 years ago] possède un contour prosodique différent dans l'exemple suivant en fonction de la question en discussion et non pas simplement en fonction de l'opposition topic/-focus.

Ces questions font partie intégrante de la thèse en cours de Fanny Krimou (en co-encadrement avec F. Nemo), je ne développerai pas plus loin cette problématique qui cependant a recours aux mêmes outils d'annotation.

(7) John went to Norway three years ago.

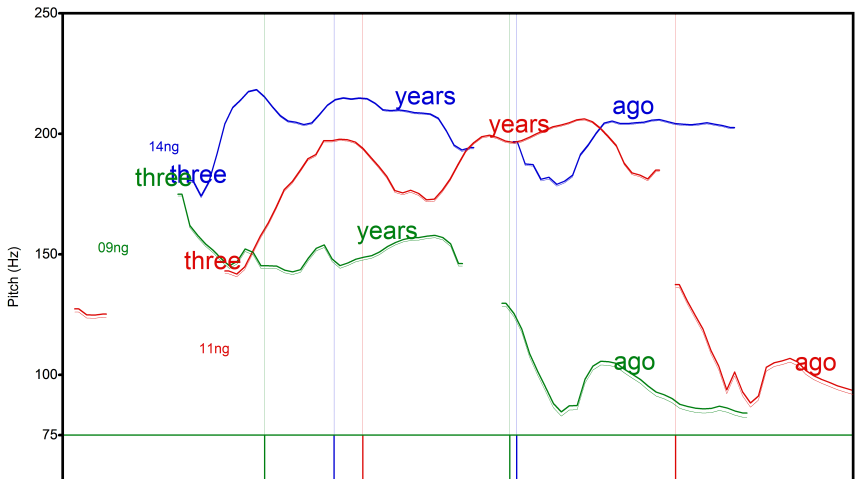


FIGURE 4.11 – Les différents patrons prosodiques pour '3 years ago'

Ces travaux sont poursuivis dans le cadre d'un projet APR-IA appelé RAVIOLI (Reconnaissance Automatique des Valeurs Injonctives à l'Oral, Langue en Interaction, porteur : L. Abouda). Ma contribution à ce projet porte essentiellement sur le design des procédures d'annotation. L'idée est d'établir le lien entre

l'interprétation pragmatique, la syntaxe et la prosodie. Le projet se concentre sur les injonctifs pragmatiques.

Bien entendu, adapter les méthodes utilisées pour le français aux langues créoles un objectif à moyen terme.

Chapitre 5

Perspectives

Sommaire

5.1 Constitution de corpus de large taille sur les langues créoles 113

5.2 Apprentissage automatique au service de la linguistique . . 114

5.3 Métagrammaires 116

5.4 Le mot de la fin 118

Conclure serait maladroit car une Habilitation à Diriger des Recherches est un début, non une fin. Je vais donc présenter quelques pistes de recherche vers lesquelles les travaux présentés précédemment me mènent.

5.1 Constitution de corpus de large taille sur les langues créoles

Plusieurs projets qui concernent la constitution de corpus oraux sur les langues créoles sont en cours de dépôt de soumission. Je l'ai déjà mentionné, un projet de transcription des données sur le kriyol de Guinée-Bissau (données recueillies par Jean-Louis Rougé) associe le LLL et l'équipe CELGA-ILTEC (Coimbra). Un projet similaire sur le créole de la Jamaïque est en cours de montage (avec Silvia

Kouwenberg), et un projet de recueil sur la Guadeloupe l'est également. Tout cela devrait s'incarner dans un projet ANR. Notre but est de permettre la constitution d'un dataset (jeu de données) qui puisse servir aux spécialistes de traitement de la parole comme terrain de développement d'outils intégrant les 'spécificités' des discours spontanés en langues créoles (dont les phénomènes d'alternance codique notamment). L'utilisation de nouvelles techniques pour le recueil et le traitement des données est à la base de ce projet.

En particulier, le recueil des données nouvelles sera effectué avec LIG-AIKUMA Gauthier et al. (2016). Nous espérons ainsi pouvoir faire passer un cap quantitatif et qualitatif aux études créoles.

5.2 Apprentissage automatique au service de la linguistique

Les techniques de Machine Learning (ML) sont à la mode. On entend même parfois dans les conférences de TAL que les linguistes sont devenus inutiles car :

— grâce au crowdsourcing (production participative) le recueil des données se fait sans les linguistes,

— grâce au ML, il est possible de se passer des analyses des linguistes (apprentissage non supervisé).

Si ce type de discours entendu à TALN peut impressionner des étudiants naïfs, et parfois même des organismes financeurs, on peut aussi en sourire. Avec un peu de recul, il semblerait même que les informaticiens TAListes sans formation en linguistique soient les plus en péril face à ces nouvelles technologies. En effet, les efforts faits par les développeurs d'outils en ML font que ces outils sont assez facilement utilisables. Les librairies ScikitLearn (Pedregosa et al. (2011)) et TensorFlow (Girija (2016)), par exemple, sont particulièrement bien documentées et un travail important est fait pour les rendre accessibles.

J'ai la chance de donner quelques heures de cours dans la Graduate School Orléans Numérique et je vois bien la facilité avec laquelle les (bons) étudiants

sont capables de s'emparer de ces outils. Comprendre leur utilisation sur un problème complexe reste une tâche ardue, mais qui mieux qu'un linguiste est capable de comprendre la complexité d'un problème linguistique ? Il reste bien entendu des linguistes réfractaires à toute utilisation des nouvelles technologies, et c'est d'ailleurs très bien ainsi, mais je peux constater, par l'exemple de mon équipe, l'évolution des chercheurs vers l'utilisation autonome d'outils toujours plus puissants. Je me souviens du temps où les linguistes n'avaient aucune autonomie dans l'usage des outils et dépendaient totalement du bon vouloir des ingénieurs d'études pour leurs expérimentations. Ce temps est révolu.

Il me semble que le ML peut être un bon auxiliaire au travail du linguiste lorsqu'il permet de réduire le temps de traitement de certains problèmes 'simples' qui se posent aux linguistes. Je pense en particulier au problème de la transcription des enregistrements de terrain. Le logiciel persephone Adams et al. (2018) est un exemple stimulant d'utilisation du ML afin d'aider le linguiste. On trouve, sur des niveaux d'analyse plus élevés (syntaxe et discours notamment) des méthodes intéressantes, comme le cycle MATTER (Model-Annotate-Train-Test-Evaluate-Revise, cf. Ide and Pustejovsky (2017)). Ceci est très stimulant car ces techniques permettent un changement d'échelle formidable dans le nombre de données disponibles pour l'analyse linguistique.

Tout cela contribue probablement plus à la disparition de l'informaticien TA-Liste "modèle 1990" qu'à la disparition du linguiste expert.

Cependant, il convient de rester prudent face à la fascination pour les solutions techniques aux problèmes complexes posés par les langues naturelles. Les solutions techniques ne se substituent pas à la connaissance du problème. Et ici, la linguistique de terrain a de beaux jours devant elle.

On remarquera que les techniques de ML supervisées nécessitent un jeu de données annotées vaste et de qualité. Ceci a un prix. Tout le monde n'a pas la puissance technique et financière permettant d'utiliser (ou de rivaliser avec) Amazon Mechanical Turk. D'un autre côté, les techniques non supervisées (par exemple : Godard et al. (2017)) restent encore de bas niveau (segmentation et alignement de

mots) et demandent toujours la validation des résultats sur un corpus 'gold' réalisé par des linguistes experts.

Dans ce contexte, les méthodes 'symboliques' restent toujours d'actualité.

5.3 Métagrammaires

Pour les langues peu dotées (en ressources linguistiques), les outils d'analyse tels que XMG2 (v. le chapitre Structure) me semblent très intéressants. Le travail que j'ai effectué sur le créole m'incite à envisager des développements de cet outil vers d'autres 'dimensions'. XMG propose une dimension (à comprendre comme un module) syntaxique, une dimension sémantique et une dimension morphologique. Je pense qu'il est nécessaire d'ajouter une dimension phonologique (ou morpho-phonologique) permettant traiter de façon satisfaisante plusieurs problèmes auxquels j'ai été confronté. Dans Duchier et al. (2012a), nous avons eu recours à un post-traitement morpho-phonologique de façon à ajuster la forme phonologique du mot à sa composition morphologique. Ceci n'est cependant pas satisfaisant. Je souhaiterais qu'un jeu de contraintes sur un plan distinct (phonologique) puisse mener au choix de la forme phonologique correcte sans passer par un post-traitement. Par ailleurs, les contraintes sur les allomorphes ne devraient pas figurer dans la dimension syntaxique. Prenons un exemple en créole haïtien. On trouve plusieurs allomorphes du défini (spécifique) *a, la, an, lan, nan*.

(1) a. loto -a
voiture DEF
'la voiture'

b. chimen -an
chemin DEF
'le chemin'

c. tab -la
table DEF
'la table'

d. bank -lan
banque DEF
'la banque'

Comme on le constate, le choix est conditionné phonologiquement. Résoudre ce problème par un trait phonologique présent dans la dimension syntaxique n'est pas satisfaisant. Par ailleurs, les données montrent que la contrainte sur le choix de l'allomorphe est liée à la forme du mot qui précède, indépendamment de son statut syntaxique :

(2) chimen an prenn nan
chemin je prendre DEF
'le chemin que j'ai pris'

Ceci plaide pour un traitement phonologique (ou morpho-phonologique) indépendant de la syntaxe.

Cette dimension reste à définir. Des contacts avec Simon Petitjean ont été pris dans ce sens et on ne peut qu'espérer que ce projet se concrétise.

La description d'autres langues (peu dotées notamment) avec XMG me paraît être une piste de recherche stimulante. J'ai vu la capacité de mes étudiants à s'emparer de cet outil pour décrire des phénomènes syntaxiques [1] et je suis convaincu que la description de langues peu dotées pourrait gagner en qualité grâce à XMG.

Mais une autre piste de développement liée aux métagrammaires me tient à coeur. Celle-ci repose sur l'idée que les grammaires codées par des linguistes, surtout s'ils sont comme moi, ne sont pas optimales. Je retiens de discussions avec Timm Lichte (HHU, Düsseldorf) l'idée que l'on pourrait chercher à obtenir des règles à partir de code optimisé. La grammaire que je décris dans le chapitre Structure est un bon exemple de compromis entre un besoin de généralisation et un besoin de lisibilité. A partir d'un certain niveau d'abstraction, les règles (et les fragments XMG) deviennent difficilement lisibles par un linguiste. Elles deviennent alors difficilement modifiables. Il serait intéressant de pouvoir optimiser

1. Il s'agit de stages de L3 et de projets tutorés.

le code et de déduire automatiquement des généralisations qui échappent au linguiste.

5.4 Le mot de la fin

Dans ce mémoire, j'ai cherché à présenter mes travaux de façon aussi cohérente que possible. Il va de soi qu'il s'agit d'une construction a posteriori. Le hasard des rencontres et des opportunités ainsi que certains contextes favorables influent sur le cours des recherches. La recherche est une activité humaine. La linguistique de terrain dépend encore plus encore que d'autres branches de la linguistique des contingences matérielles et sociales. En particulier, il devient de plus en plus difficile de faire des recherches de terrain en Afrique subsaharienne, là précisémment où se trouvent encore de nombreuses langues peu décrites, et qui sont justement d'une grande importance pour l'étude des origines des langues créoles. L'accueil de doctorants n'est pas très aisé et les échanges avec les collègues du Sud sont parfois difficiles, tant leurs conditions de travail sont parfois précaires. Dans ce contexte, attirer des doctorants vers la linguistique de terrain et la créolistique n'est pas une tâche aisée. Les équipes de recherches ont leurs contraintes également : il faut publier, avoir une masse critique pour être 'visible' des institutions...

Tout cela pourrait être un frein aux projets en cours et à ceux que j'ai annoncés précédemment. Pourtant, je pense que la linguistique de terrain moderne a un grand avenir devant elle. Ma participation au GDRI Créole et au GDR LIFT me permettent d'être optimiste malgré tout. Je ne crois pas que les avancées majeures en linguistique viendront de n$^{\text{ièmes}}$ travaux supplémentaires sur le français ou l'anglais, ni même de l'augmentation de la taille des corpus par deux, quatre ou dix sur ces langues. Tant de langues restent à décrire, partiellement ou presque totalement, et c'est certainement là qu'il faut accentuer les recherches. C'est en tout cas la direction qu'indique ma boussole.

Bibliographie

Abeillé, Anne. 1993. *Les Nouvelles Syntaxes. Grammaires d'unification et analyse du français*. Armand Colin, Paris.

Abeillé, Anne. 2002. *Une Grammaire électronique du Français*. CNRS Editions, Paris.

Abeillé, Anne. 2011. The French treebank : applications and extensions. In *[s.t.]*, Stuttgart, Unknown Region.

Abeillé, Anne and Danièle Godard. 2007. Les relatives sans pronom relatif. In M. Abecassis ;, editor, *Le francais parlé, Normes et variations*. L'Harmattan, pages 37–60.

Abeillé, Anne and Owen Rambow. 2000. *Tree Adjoining Grammars : Formalisms, Linguistic Analysis, and Processing*. CSLI publications.

Aboh, Enoch O. 2016. Creole distinctiveness. *Journal of Pidgin and Creole Languages*, 31(2) :400–418.

Aboh, Enoch Oladé. 2015. *The Emergence of Hybrid Grammars : Language Contact and Change*. Cambridge University Press.

Adams, Oliver, Trevor Cohn, Graham Neubig, Hilaria Cruz, Steven Bird, and Alexis Michaud. 2018. Evaluating phonemic transcription of low-resource tonal languages for language documentation. In *Proceedings of LREC 2018*.

Antoine, Jean-Yves, Anaïs Lefeuvre, and Emmanuel Schang. 2016. Codage en chaîne ou en première mention de la coréférence : Approcher la structure des chaînes de référence par comparaison des deux annotations. *SHS Web of Conferences*, 27(nil) :02001.

Antoine, Jean-Yves, Jakub Waszczuk, Anaïs Lefeuvre-Halftermeyer, Lotfi Abouda, Emmanuel Schang, and Agata Savary. 2017. Temporal@ODIL Project : Adapting ISO-TimeML to Syntactic Treebanks for the Temporal Annotation of Spoken Speech. In *Thirteenth Joint ACL - ISO Workshop on Interoperable Semantic Annotation (ISA-13), 12th International Conference on Computational Semantics (IWCS'2017), Montpellier, France.*, Montpellier, France.

Bakker, Peter, Finn Borchsenius, Carsten Levisen, and Eeva Sippola. 2017. *Creole studies–phylogenetic approaches*. John Benjamins Publishing Company.

Bandeira, Manuele. 2017. *Reconstrução fonológica e lexical do protocrioulo do Golfo da Guiné*. Ph.D. thesis, Universidade de São Paulo.

Baptista, Marlyse and Jacqueline Guéron. 2007. *Noun phrases in creole languages : a multi-faceted approach*, volume 31. John Benjamins Publishing.

Bhatt, Parth and Tjerk Hagemeijer. Complex onsets in santome : phonological innovation in creoles ?

Bickerton, Derek. 1981. Roots of language.

Bittar, André, Pascal Amsili, Pascal Denis, and Laurence Danlos. 2011. French timebank : an iso-timeml annotated reference corpus. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies : short papers-Volume 2*, pages 130–134, Association for Computational Linguistics.

Blackburn, Patrick. 2005. Representation and inference for natural language : A first course in computational semantics.

Bloomfield, Leonard. 1962. Language. 1933. *Holt, New York*.

Bouchard, Denis. 2003. Les sn sans déterminant en français et en anglais. *Essais sur la grammaire comparée du français et de l'anglais*, pages 55–95.

Bouquiaux, Luc and Jacqueline MC Thomas. 1976. *Enquête et description des langues à tradition orale*, volume 1. Peeters Publishers.

Brassier, Maëlle, Alexis Puret, Augustin Voisin-Marras, and Loïc Grobol. 2018. Classification par paires de mention pour la résolution des coréférences en français parlé interactif. In *Conférence jointe CORIA-TALN-RJC 2018*, ATALA and ARIA, Rennes, France.

Candito, Marie-Hélène. 1999. *Organisation modulaire et paramétrable de grammaires électroniques lexicalisées*. Ph.D. thesis.

Chaudenson, Robert. 2004. *La créolisation : théorie, applications, implications*. Editions L'Harmattan.

Citko, Barbara. 2014. *Phase theory : An introduction*. Cambridge University Press.

Coelho, Margarida, Cíntia Alves Valentina Coia, Donata Luiselli, Antonella Useli, Tjerk Hagemeijer, António Amorim, Giovanni Destro-Bisol, and Jorge Rocha. 2008. Human microevolution and the atlantic slave trade : a case study from sao tome. *Current Anthropology*, 49(1) :134–143.

Colot, Serge and Ralph Ludwig. 2013. Guadeloupean Creole structure dataset. In Susanne Maria Michaelis, Philippe Maurer, Martin Haspelmath, and Magnus Huber, editors, *Atlas of Pidgin and Creole Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Cosme, Abigail Tiny. 2014. *As relações filogenéticas entre os crioulos do Golfo da Guiné*. Ph.D. thesis.

Crabbé, Benoit. 2005. *Représentation informatique de grammaires d'arbres fortement lexicalisées : le cas de la grammaire d'arbres adjoints*. Ph.D. thesis, Université Nancy 2.

Creissels, Denis. 1991. *Description des langues négro-africaines et théorie syntaxique*. Ellug.

Creissels, Denis. 1994. *Aperçu sur les structures phonologiques des langues négro-africaines*. Ellug.

Creissels, Denis. 1995. *Eléments de syntaxe générale*. Presses Universitaires de France-PUF.

Creissels, Denis. 2006. *Syntaxe générale : une introduction typologique*. Hermes science.

De Marneffe, Marie-Catherine, Marta Recasens, and Christopher Potts. 2015. Modeling the lifespan of discourse entities with application to coreference resolution. *Journal of Artificial Intelligence Research*, 52 :445–475.

DeCamp, David. 1971. *Introduction : The study of pidgin and creole languages*. Cambridge University Press.

Déprez, Viviane. 2005. Morphological number, semantic number and bare nouns. *Lingua*, 115(6) :857–883.

Déprez, Viviane and Fabiola Henri. 2018. *Negation and Negative Concord : The view from Creoles*. Contact Language Library. John Benjamins Publishing Company.

Duchier, Denys, Brunelle Magnana Ekoukou, Yannick Parmentier, Simon Petitjean, Emmanuel Schang, and others. 2012a. Décrire la morphologie des verbes en ikota au moyen d'une métagrammaire. *JEP-TALN-RECITAL 2012*, page 97.

Duchier, Denys, Brunelle Magnana Ekoukou, Yannick Parmentier, Simon Petitjean, Emmanuel Schang, and others. 2012b. Describing morphologically-rich languages using metagrammars : A look at verbs in Ikota. *Language Technology for Normalisation of Less-Resourced Languages*, page 55.

Duchier, Denys, Yannick Parmentier, Simon Petitjean, and Emmanuel Schang. 2017. Produire des ressources électroniques à partir de descriptions formelles : application aux langues peu dotées. In *DiLiTAL-Diversité Linguistique et TAL*, pages 24–32.

Ferraz, Luiz Ivens. 1978. The creole of São Tomé. *African studies*, 37(1) :3–68.

Ferraz, Luiz Ivens. 1987. The liquid in the Gulf of Guinea Creoles. *African Studies*, 46(2) :287–295.

Frank, Robert. 2002. *Phrase Structure Composition and Syntactic Dependencies*. MIT Press, Cambridge, Mass.

Frank, Robert. 2006. Phase theory and tree adjoining grammar. volume 116. Lingua, Elsevier, pages 145–202.

Gadelii, Karl. 2007. The bare np in lesser antillean. *CREOLE LANGUAGE LIBRARY*, 31 :243.

Gauthier, Elodie, David Blachon, Laurent Besacier, Guy-Noel Kouarata, Martine Adda-Decker, Annie Rialland, Gilles Adda, and Grégoire Bachman. 2016. Lig-aikuma : A mobile app to collect parallel speech for under-resourced language studies. In *Interspeech 2016 (short demo paper)*.

Girija, Sanjay Surendranath. 2016. Tensorflow : Large-scale machine learning on heterogeneous distributed systems. *Software available from tensorflow. org*.

Glaude, Herby. 2013. Corpus Créoloral. oai :crdo.vjf.cnrs.fr :crdo-GCF, SFL Université Paris 8 - LLL Université Orléans.

Godard, Pierre, Gilles Adda, Martine Adda-Decker, Juan Benjumea, Laurent Besacier, Jamison Cooper-Leavitt, Guy-Noël Kouarata, Lori Lamel, Hélène Maynard, Markus Müller, et al. 2017. A very low resource language speech corpus for computational language documentation experiments. *arXiv preprint arXiv :1710.03501*.

Grimshaw, J. 1991. Extended projection. *Ms., Brandeis University*.

Grimshaw, Jane. 2000. Locality and extended projection. *Amsterdam Studies in the Theory and History of Linguistic Science Series 4*, pages 115–134.

Grobol, Loïc. 2019. Neural Coreference Resolution with Limited Lexical Context and Explicit Mention Detection for Oral French. In *Second Workshop on Computational Models of Reference, Anaphora and Coreference (CRAC19)*, Minneapolis, United States.

Gundel, Jeanette K. 2003. Information structure and referential givenness/newness : How much belongs in the grammar. In *Proceedings of the HPSG'03 Conference*, pages 143–162.

Gundel, Jeanette K. 2010. Reference and accessibility from a Givenness Hierarchy perspective. *International Review of Pragmatics*, 2(2) :148–168.

Gundel, Jeanette K, Nancy Hedberg, and Ron Zacharski. 2001. Definite descriptions and cognitive status in English : Why accommodation is unnecessary. *English Language and Linguistics*, 5(02) :273–295.

Gundel, Jeanette K., Michael Hegarty, and Kaja Borthen. 2003. Cognitive status, information structure, and pronominal reference to clausally introduced entities. *Journal of Logic, Language and Information*, 12(3) :281–299.

Hagège, Claude. 1996. *L'homme de paroles : contribution linguistique aux sciences humaines*. Fayard.

Hagemeijer, Tjerk. 2007. *Clause structure in Santome*. Ph.D. thesis, University of Lisbon.

Hagemeijer, Tjerk. 2011. The gulf of guinea creoles : genetic and typological relations. *Journal of Pidgin and Creole Languages*, 26(1) :111–154.

Hassamal, Shrita. 2017. *Grammar of Mauritian adverbs*. Ph.D. thesis, Sorbonne Paris Cité.

Hazaël-Massieux, Guy. 1996. Les créoles(problèmes de genèse et de description).

Hazaël-Massieux, Marie-Christine. 2008. *Textes anciens en créole français de la Caraïbe : Histoire et analyse*. Éditions Publibook.

Henri, Fabiola. 2010. *A Constraint-Based Approach to verbal constructions in Mauritian*. Ph.D. thesis, Paris Diderot.

Henri, Fabiola and Anne Abeillé. 2007. The syntax of copular constructions in Mauritian. In *14th Conference on HPSG*, pages 130–149, CSLI Publications, Stanford, United States.

Hinzen, Wolfram. 2006. *Mind design and minimal syntax*. Oxford University Press.

Holm, John. 1989. Pidgins and creoles. vol. 1 & 2. *Cambridge : CUP*.

Homer, Vincent. 2013. On the Nonexistence of Negative Quantifiers : The Case of Haitian Creole.

Ide, Nancy and James Pustejovsky. 2017. *Handbook of linguistic annotation*. Springer.

Jespersen, Otto. 2013. *Language : its nature and development*. Routledge.

Kaplan, Ronald M, Joan Bresnan, et al. 1982. Lexical-functional grammar : A formal system for grammatical representation. *Formal Issues in Lexical-Functional Grammar*, 47 :29–130.

de La Clergerie, Éric Villemonte. 2010. Convertir des dérivations tag en dépendances. In *17e Conférence sur le Traitement Automatique des Langues Naturelles-TALN 2010*.

Labov, William. 1971. The notion of 'system'in creole studies. *Pidginization and creolization of languages*, 447 :472.

Landragin, Frédéric. 2017. Analyse, visualisation et identification automatique des chaînes de coréférences : des questions interdépendantes ? *Langue française*, (3) :17–34.

Lefeuvre, Anaïs, Jean-Yves Antoine, Agata Savary, Emmanuel Schang, Lotfi Abouda, Denis Maurel, and Iris Eshkol. 2014. Annotation de la temporalité en corpus : contribution à l'amélioration de la norme TimeML. In *Proceedings of TALN 2014*.

Lefeuvre, Anais, Jean-Yves Antoine, and Emmanuel Schang. 2014. Le corpus ANCOR_Centre et son outil de requêtage : application à l'étude de l'accord en genre et nombre dans les coréférences et anaphores en français parlé. In *SHS Web of Conferences*, volume 8, pages 2691–2706, EDP Sciences.

Lefeuvre-Halftermeyer, Anaïs, Jean-Yves Antoine, Alain Couillault, Emmanuel Schang, Lotfi Abouda, Agata Savary, Denis Maurel, Iris Eshkol-Taravella, and Delphine Battistelli. 2016. Covering various Needs in Temporal Annotation : a Proposal of Extension of ISO TimeML that Preserves Upward Compatibility. In *LREC 2016*, Portorož , Slovenia.

López, Luis, Artemis Alexiadou, and Tonjes Veenstra. 2017. Code-switching by phase. *Languages*, 2(3) :9.

Lorenzino, Gerardo. 1998. *The Angolar Creole Portuguese of São Tomé : its grammar and sociolinguistic history*. Ph.D. thesis, City University of New York.

Magnana Ekoukou, Brunelle. 2015. *Description de l'Ikota (B25), langue bantu du Gabon. Implémentation de la morphosyntaxe et de la syntaxe*. Ph.D. thesis, Orléans.

Mangeot, Mathieu and Emmanuel Schang. 2018. *TALAf 2018 : Traitement Automatique des Langues Africaines*.

Manuélian, Hélène. 2003. *Descriptions définies et démonstratives : analyses de corpus pour la génération de textes*. Ph.D. thesis, Université de Nancy 2.

Maurer, Philippe. 1995. *L'angolar : un créole afro-portugais parlé à São Tomé : notes de grammaire, textes, vocabulaires*, volume 16. Buske Verlag.

Maurer, Philippe. 2008. A first step towards the analysis of tone in Santomense. *Roots of Creole Structures*, pages 253–261.

McCrindle, Karen Lyda. 1999. Temps, mode et aspect, les creoles des Caraibes a base lexicale française.

McWhorter, John. 2001. The world's simplest grammars are creole grammars. *Linguistic typology*, 5(2) :125–66.

McWhorter, John H. 2011. *Linguistic simplicity and complexity : Why do languages undress ?*, volume 1. Walter de Gruyter.

Meijer, Guus and Pieter Muysken. 1977. On the beginnings of pidgin and creole studies : Schuchardt and hesseling. In Albert Valdman, editor, *Pidgin and creole linguistics*. Bloomington : Indiana University Press.

Michaelis, Susanne Maria, Philippe Maurer, Martin Haspelmath, and Magnus Huber, editors. 2013. *APiCS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Millour, Alice and Karën Fort. 2018. Krik : First steps into crowdsourcing pos tags for kréyòl gwadloupéyen. In *Proceedings of the Eleventh International*

Conference on Language Resources and Evaluation (LREC 2018), European Language Resources Association (ELRA), Paris, France.

Morel, Michel, Anne Lacheret-Dujour, Chantal Lyche, François Poiré, and others. 2006. Vous avez dit proéminence ? *Actes des XXVIes journées d'étude sur la parole*, pages 183–186.

Mufwene, Salikoko. 2001a. Les continua créoles, linguistiques, et langagiers.

Mufwene, Salikoko S. 1997. Introduction : Understanding speech continua. *World Englishes*, 16(2) :181–184.

Mufwene, Salikoko S. 2001b. *The ecology of language evolution*. Cambridge University Press.

Mufwene, Salikoko S. 2005. *Créoles, écologie sociale, évolution linguistique*. Editions L'Harmattan.

Mühlhäusler, Peter. 1986. *Pidgin and creole linguistics*. Blackwell Oxford.

Muysken, PC, B Jansen, and H Koopman. 1978. Serial verbs in the creole languages.

Muzerelle, Judith, Emmanuel Schang, Jean-Yves Antoine, Iris Eshkol, Denis Maurel, Aurore Boyer, and Damien Nouvel. 2012. Annotations en chaînes de coréférences et anaphores dans un corpus de discours spontané en français. *SHS Web of Conferences*, 1 :2497–2516.

Obin, Nicolas, Julie Beliao, Christophe Veaux, and Anne Lacheret. 2014. Slam : Automatic stylization and labelling of speech melody. In *Speech Prosody*, pages 246–250.

Ogrodniczuk, Maciej, Katarzyna Głowinska, Mateusz Kopec, Agata Savary, and Magdalena Zawisławska. 2013. Polish coreference corpus. *Journalism*, 3(7,078) :19–53.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn : Machine learning in Python. *Journal of Machine Learning Research*, 12 :2825–2830.

Petitjean, Simon. 2014. *Génération modulaire de grammaires formelles*. Ph.D. thesis, Université d'Orléans.

Petitjean, Simon, Denys Duchier, and Yannick Parmentier. 2016. Xmg 2 : describing description languages. In *International Conference on Logical Aspects of Computational Linguistics*, pages 255–272, Springer.

Petitjean, Simon and Emmanuel Schang. 2018. Sentential negation and negative words in guadeloupean creole.

Pfänder, Stefan. 2000. *Aspekt und Tempus im Frankokreol*. G. Narr.

Pinker, Steven. 1995. *The language instinct : The new science of language and mind*, volume 7529. Penguin UK.

Pinker, Steven. 1999. *L'instinct du langage*. Odile Jacob.

Pollard, Carl and Ivan A Sag. 1994. *Head-driven phrase structure grammar*. University of Chicago Press.

Prudent, Lambert-Félix. 1999. *Des baragouins à la langue antillaise : analyse historique et sociolinguistique du discours sur le créole*. Editions L'Harmattan.

Radford, A. 1988. *Transformational grammar : A first course*. Cambridge Univ Pr.

Radford, A. 1997. *Syntax : a minimalist introduction*. Cambridge Univ Pr.

Reboul, Anne and Jacques Moeschler. 1998. Pragmatique du discours. *De l'interprétation de l'énoncé à l'interprétation du discours. Paris, Armand Colin*.

Recasens, Marta, Eduard Hovy, and M. Antònia Martí. 2011. Identity, non-identity, and near-identity : Addressing the complexity of coreference. *Lingua*, 121(6) :1138–1152.

Recasens, Marta, Marie-Catherine de Marneffe, and Christopher Potts. 2013. The life and death of discourse entities : Identifying singleton mentions. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, pages 627–633.

Romaine, Suzanne. 2017. *Pidgin and creole languages*. Routledge.

Rougé, Jean-Louis. 1992. Les langues des tonga. *Actas do colóquio sobre crioulos de base lexical portuguesa*, pages 171–176.

Rougé, Jean-Louis. 2004. *Dictionnaire étymologique des créoles portugais d'Afrique*. Karthala.

Rougé, Jean-Louis and Emmanuel Schang. 2006. The origin of the liquid consonant in Saotomense Creole. *The Structure of Creole Words. Segmental Syllabic and Morphological Aspects*, pages 25–37.

Rougé, Jean-Louis and Emmanuel Schang. 2012. Classificar, denominar as línguas de São Tomé.

Rougé, Jean-Louis and Emmanuel Schang. 2012. Histoire des créoles et génétique : le cas de l'angolar. *Sciences & Techniques du Langage, Dakar*.

Rougé, Jean-Louis and Emmanuel Schang. 2012. Post creolisation evolution : the case of santomense. Presented at the 9 th Creolistics Workshop, Aarhus university.

Rougé, Jean-Louis and Emmanuel Schang. 2013. Ce qu'enseigne la comparaison des créoles portugais d'afrique. In *Actes del 26é Congrés de Lingüística i Filologia Romàniques (València, 6-11 de setembre de 2010)*, pages 629–639.

Russell, Bertrand. 1905. On denoting. definite descriptions : A reader, ed. by gary ostertag, 35–49.

dos Santos Agostinho, Ana Lívia. 2016. *Fonologia do lung'Ie*. Lincom GmbH.

Schang, Emmanuel. 2000. *L'émergence des créoles portugais du Golfe de Guinée*. Ph.D. thesis, Nancy 2.

Schang, Emmanuel. 2002. *L'émergence des créoles portuguais du golfe de Guinée : thèse pour obtenir le grade de docteur de l'université Nancy 2 en sciences du langage*. Presses universitaires du septentrion.

Schang, Emmanuel. 2003. Syllable structure and creolization in Saotomense. *Plag (ed.)*, pages 109–120.

Schang, Emmanuel. 2004. Identification des langues : Que faire des créoles ? In *MIDL04*, pages 19–23, ENST.

Schang, Emmanuel. 2013. Extended Projections in a Guadeloupean TAG Grammar. In *Workshop on High-level Methodologies for Grammar Engineering@ ESSLLI 2013*, page 49.

Schang, Emmanuel. 2018. A metagrammatical approach to periphrasis in gwadloupéyen. *Quaderni di Linguistica e Studi Orientali*, 4 :131–149.

Schang, Emmanuel, Jean-Yves Antoine, and Anaïs Lefeuvre-Halftermeyer. 2017. Les chaînes coréférentielles en créole de la Guadeloupe. In *TALN'2017, atelier DILITAL*, Orléans, France.

Schang, Emmanuel, Denys Duchier, Brunelle Magnana Ekoukou, Yannick Parmentier, and Simon Petitjean. 2012a. Describing São Tomense Using a Tree-Adjoining Meta-Grammar. In *11th International Workshop on Tree Adjoining Grammars and Related Formalisms (TAG+ 11)*.

Schang, Emmanuel, Denys Duchier, Brunelle Magnana Ekoukou, Yannick Parmentier, and Simon Petitjean. 2012b. Describing Sao Tomense Using a Tree-Adjoining Meta-Grammar. In *11th International Workshop on Tree Adjoining Grammars and Related Formalisms (TAG+ 11)*, pages 82–89.

Schnedecker, Catherine and Frédéric Landragin. 2014. Les chaînes de référence : présentation. *Langages*, 195(3) :3.

Schuchardt, Hugo. 1914. The language of the saramaccans. *Thomas L. Markey (ed. & trans.), The ethnography*.

Sing, Guillaume Fon. 2017. Creoles are not typologically distinct from non-creoles. *Language Ecology*, 1(1) :44–74.

Stewart, William A. 1969. Urban negro speech : Sociolinguistic factors affecting english teaching. *Florida FL Rep*.

Tabouret-Keller, Andrée. 1988. Contacts de langues : Deux modèles du xixème siècle et leurs rejetons aujourd'hui. *Langage et société*, 43(1) :9–22.

Tabouret-Keller, Andrée. 1997. *Les enjeux de la nomination des langues*, volume 95. Peeters Publishers.

Vaillant, Pascal. 2008a. Grammaires factorisées pour des dialectes apparentés. In *TALN 2008 : Actes de la 15ème conférence annuelle sur le Traitement Automatique des Langues Naturelles*, pages p. 159–168, ATALA (Association pour le Traitement Automatique des Langues), Avignon, France. 10 pages. Actes de la 15eme conference annuelle sur le Traitement Automatique des Langues Naturelles (TALN 2008), Avignon, France, 9-13 juin 2008.

Vaillant, Pascal. 2008b. Grammaires factorisées pour des dialectes apparentés. In *TALN 2008 : Actes de la 15ème conférence annuelle sur le Traitement Automatique des Langues Naturelles*.

Valdman, Albert, editor. 1977. *Pidgin and creole linguistics*. Indiana University Press Bloomington.

Veenstra, Tonjes. 2009. Verb allomorphy and the syntax of phases. *Complex processes in new languages*, pages 99–113.

Velupillai, Viveka. 2015. *Pidgins, creoles and mixed languages : An introduction*, volume 48. John Benjamins Publishing Company.

Walker, Marilyn A, Aravind Krishna Joshi, and Ellen Friedman Prince. 1998. *Centering theory in discourse*. Oxford University Press.

Widlöcher, Antoine and Yann Mathet. 2012. The Glozz platform : a corpus annotation and mining tool. In *Proceedings of the 2012 ACM symposium on Document engineering*, pages 171–180, ACM.

Winford, Donald. 2012. Creole Languages. *The Oxford Handbook of Tense and Aspect*.

Xia, Fei. 2001. *Automatic grammar generation from two different perspectives*. University of pennsylvania.

Zribi-Hertz, A. 1996. *L'anaphore et les pronoms*. Presses Universitaires du Septentrion.

Zribi-Hertz, Anne. 2011. Pour un modèle diglossique de description du français : quelques implications théoriques, didactiques et méthodologiques. *Journal of French Language Studies*, 21(2) :231–256.

Zribi-Hertz, Anne and Loïc Jean-Louis. 2013. From noun to name : definiteness marking in modern martinikè. *Crosslinguistic studies on Noun Phrase structure and reference*, pages 269–315.