Développement de ressources langagières et d'outils de TAL pour le créole mauricien

Mémoire de Master 1 Langue et Informatique

présenté par

Harmonie Begue

Sous la direction de K. Fort et A. Millour

16 septembre 2019

Remerciements

Je remercie tout d'abord Karën Fort qui a bien voulu m'encadrer pour ce travail et qui m'a donné l'occasion de d'interagir, lors du court instant qu'est la durée de ce projet, avec des chercheur·e·s passionné·e·s et inspirant·e·s.

Un grand merci à Alice Millour, pour sa disponibilité, ses conseils pertinents, ses encouragements – qui m'ont permis de garder le cap jusqu'ici.

Merci également à Gaël Lejeune de m'avoir inspirée pour mener ce travail en me faisant découvrir les travaux menés par Alice Millour.

Je remercie Fabiola Henri d'avoir partagé avec moi son expertise du *kreol morisien* et d'avoir pris le temps de corriger mes annotations manuelles.

Merci à Shrita Hassamal, Mushina Alleesaib, Jimmy Harmon et Arnaud Carpooran qui ont bien voulu me rencontrer lors de mon passage à Maurice, et qui ont aidé à la diffusion de la plateforme.

Je ne remercierai jamais assez ma mère, Marlène, qui a bien voulu me relire et dont le soutien a été sans faille malgré la distance.

Je remercie particulièrement Islam et Célia pour leurs inlassables encouragements et d'avoir toujours su m'apaiser lors de mes moments de doute.

Merci également à Jean, d'être resté à mes côtés et d'avoir fait preuve de tant de patience à mon égard durant cette période difficile.

Enfin, merci à tous les répondants au sondage et aux participants à la plateforme sans qui rien n'aurait abouti.

Table des matières

Table des figures

Liste des tableaux

Introduction

Le créole mauricien (*kreol morisien*, dorénavant KM) est considéré comme une langue « peu dotée » dans le domaine du TAL, et il a un statut assez particulier. Effectivement, le KM n'avait qu'un statut de langue orale et ce n'est que récemment, depuis une vingtaine d'années, qu'un mouvement peut être observé vers la normalisation et la diffusion d'une graphie, d'une grammaire et d'une orthographe de cette langue.

Cela pourrait expliquer la situation du KM dans un contexte de TAL aujourd'hui, où peu de travaux ont été menés à l'heure actuelle. Au premier abord, cela semble compliquer son informatisation et son traitement automatique à cause de la variété de formes lexicales qui pourraient être rencontrées pour une seule occurrence, mais avant tout à cause du manque de ressources et d'outils linguistiques accessibles capables de gérer cela.

La création de ressources langagières et technologies du TAL et leur évaluation semble primordiale pour le KM qui est en grande partie exclu du monde numérique.

Ce mémoire vise donc à dévélopper des ressources et des outils de TAL pour le créole mauricien afin que sa visibilité en ligne soit amplifiée et que ses locuteurs soient plus confiants à partager leurs productions écrites dans cette langue.

La collecte de donnée étant une étape souvent coûteuse, nous avons choisi de suivre les traces de Millour et Karèn Fort 2018 pour la collecte de corpus à travers la production participative (« crowsourcing ») – une méthodologie encore très peu appliquée aux langues peu dotées. Avec une plateforme adaptée pour le KM, nous procéderons ensuite à appliquer les mêmes méthodes suivies par Millour et Karèn Fort 2018, à savoir :

— entraîner un outil d'annotation sur un corpus de référence,
— puis l'utiliser sur les textes obtenus des contributions des locuteurs,
— et finalement, donner la possiblilité aux utilisateurs de la plateforme de corriger les annotations morphosyntaxiques produites automatiquement par l'outil.

Grâce à ce travail, nous espérons encourager les locuteurs du KM à s'exprimer dans leur langue en ligne afin de réduire le fossé entre le KM et l'anglais ou le français, très utilisés dans le numérique à Maurice, pour que toute la population se sente intégrée dans la démarche d'informatisation et de diffusion du KM dans le maximum de contextes socioculturels.

Chapitre 1

Présentation du sujet

1.1 Langues peu dotées et TAL

Les recherches qui ont eu lieu dans le domaine du traitement automatique des langues (TAL) pendant les dernières décennies s'étaient essentiellement focalisées sur le traitement des langues européennes (l'anglais, le français, l'allemand entre autres) et d'Asie de l'Est (principalement le chinois), au détriment de la multitude de langues minoritaires dans le monde. De ce fait, ces langues se retrouvent avec peu – voire aucune – ressource de traitement automatique, les excluant du phénomène de numérisation des langues.

Ces langues sont communément appelées « peu dotées », du fait de leur manque d'outils et de ressources électroniques.

Le principal enjeu d'outiller les langues peu dotées est d'éviter leur disparition au profit des langues plus vastement parlées avec lesquelles elles co-existent [1]. Effectivement, 90 % des langues qui existent encore aujourd'hui auront probablement disparues d'ici 2100 selon le linguiste Michael Krauss (1992). Comme le remarque très justement Benjamin Philip King (2015), Krauss avait émis cette estimation avant l'émergence d'Internet.

De fait, Internet pourrait desservir ces langues : le processus de disparition des langues peu dotées pourrait s'accélérer si leurs locuteurs sentent qu'ils doivent s'exprimer dans une langue plus « majoritairement » utilisée que la leur pour communiquer. Néanmoins, nous pouvons également imaginer un cas où une présence plus répandue de ces langues sur Internet pourrait être bénéfique, au contraire, à leur survie si leurs locuteurs y trouvent des outils de communication adéquats pour partager du contenu et s'exprimer sans concession sur la langue employée.

Tel est le contexte dans lequel ce projet s'incrit : notre objectif est de créer et fournir des ressources adaptées pour le KM [2], afin que des outils (outils de traitement de texte, correcteurs orthographiques, outils de traduction automatique) puissent être mis en place pour sa diffusion et son utilisation sur Internet.

1. Les créoles notamment, sont apparus pour la plupart dans un contexte bilingue ou multilingue (Nye 2008).

2. Le créole mauricien n'est pour l'instant pas une langue en voie de disparition (il n'est pas répertorié dans l'Atlas UNESCO des langues en danger dans le monde, voir : `http://www.unesco.org/languages-atlas/index.php`).

1.2 Langues peu dotées et méthodologies en TAL

1.2.1 La collecte de corpus : un défi technique

Dans tous les travaux d'apprentissage automatique, l'ensemble du processus est alimenté par des données, et dans le domaine du TAL, les données se présentent sous forme de corpus oraux ou écrits.

Sans un ensemble de données correctement organisé, l'efficacité d'outils de TAL peut être amoindris.

Alexis Palmer et Michaela Regneri (2013) identifient les principales difficultés qui sont rencontrées lors de la collecte de données :

— absence pure et simple de données
— données existantes mais inaccessibles
— format des données non-utilisable pour du traitement numérique
— données présentant des transcriptions / traductions / annotations inconsistantes

De plus, quand un corpus est mis en place, il n'est souvent pas sensible à l'évolution de la langue sur le long terme. Les données linguistiques ne sont souvent pas à jour, faute de moyens humains et financiers. De ce fait, les changements et/ou variations orthographiques ne sont pas toujours répertorié·e·s même pour les langues majoritaires. Sans surprise, cela s'avère encore plus compliqué pour les langues peu dotées, qui subissent différentes vagues de variation, n'ayant souvent pas d'orthographe officielle.

La solution pour rendre cela moins dramatique serait la mise à disposition libre des ressources à la communauté afin que les locuteurs y apportent eux-mêmes les différents changements qu'ils observent dans leur langue.

1.2.2 Méthodes précédemment appliquées

Les travaux existants en TAL sur les langues peu dotées peuvent être divisés en deux catégories (King 2015) : (1) les approches traitant un petit groupe de langues à la fois et (2) les approches appliquées à une grande variété de groupes de langues simultanément.

La première approche se focalise essentiellement sur une seule langue ou un petit groupe de langues similaires, et vise à la collecte de données textuelles ou orales et également à la création d'outils de TAL. Cette méthode semble fiable dans la qualité des ressources produites, même si elle est largement dépendante de l'expertise linguistique apportée. Les outils crées de cette sorte ne sont généralement pas applicables directement sur d'autres langues.

La deuxième approche implique l'utilisation d'un robot d'indexation (*web crawler*) pour parcourir le Web et, à l'aide de requêtes très précises, afficher des pages Web dans une langue peu dotée donnée. C'est ainsi que différents projets ont vu le jour : le projet Crúbadán (Scannell 2007) à réussi à construire un corpus pour

1 872 langues, le *Leipzig Corpora Collection* (Biemann et al. 2007) pour 124 langues, le *Human Language Project* (Abney et Bird 2010) qui vise au développement d'un format universel pour les corpus de texte annoté. Néanmoins, les données collectées dans le cadre de ces projets ne sont pas sous une licence libre et ne peuvent donc pas être distribuées, et requièrent plus de moyens technologiques et humains.

1.2.3 Notre méthode

Dans le cadre de ce projet, notre procéderons dans les pas de la méthode (1) décrite précédemment (section 1.2.2). Ce projet étant une adaptation du travail fait par Alice Millour et Karën Fort sur l'alsacien (2018) et le créole guadeloupéen (2018), nous y appliquerons les mêmes procédés méthodologiques, à savoir la collecte de corpus par production participative (ou « *crowdsouring* »).

La myriadisation, terme utilisé interchangeablement avec la production participative, est un néologisme qui apparaît notamment dans une publication de Benoıt Sagot et al. 2011 et vient « de l'idée qu'un certain nombre de tâches pouvaient être effectuées par des utilisateurs d'Internet, en utilisant les atouts propres à celui-ci, c'est-à-dire pouvoir accéder à un grand nombre de personnes, de manière quasi-instantanée, partout dans le monde. » [3].

Selon Burger-Helmchen et Pénin 2011, cette méthode peut s'appliquer à trois types de tâches : inventives, répétitives et de production de contenu et Millour et Karën Fort 2016 ajoutent qu'elle « peut prendre trois forme principales : la participation rémunérée et directe, bénévole et directe, bénévole et indirecte ».

1.2.4 Nos objectifs

Nos objectifs se présentent ainsi :

— donner une vision globale de la situation du KM qui s'inscrit dans les créoles à bases françaises,
— répertorier des travaux existants sur le KM et d'autres créoles français en TAL,
— analyser les résultats du sondage mené sur la relation des mauriciens avec leur créole,
— et enfin adapter une plateforme existante pour la collecte de données textuelles annotées en mauricien et l'évaluer.

3. Benoıt Sagot et al. 2011, p.2

Chapitre 2

État de l'art

2.1 Les créoles à base française

Selon Muysken, Smith et al. 1995, un créole est une langue qui a émergé à un moment précis dans le temps qu'il est possible de déterminer. Cette définition complète celle de Chaudenson 1979, qui stipule qu'un créole est « un système linguistique caractérisé par son histoire (colonisation), sa structure (autonomie par rapport aux systèmes dont il semble issu), son statut et sa fonction ». Les créoles ont donc assurément un statut de langue au niveau linguistique étant donné que leur but premier est la communication entre des locuteurs natifs et permanents d'une situation géographique précise. [1]

Dans le cadre de ce travail, nous nous intéresserons plus particulièrement aux créoles à base lexicale française, qui sont apparus au cours du XVII^{ème} et XVIII^{ème} siècles sous l'empire colonial français. Comme l'explique Hazaël-Massieux 2002, l'apparition des créoles est due à l'« évolution accélérée de formes régionales et populaires du français ».

Figure 2.1 – Les créoles à base lexicale française dans le monde (Michaelis et al. 2013)

1. En cela, il ne faut pas confondre les créoles et les pidgins, qui ont un caractère instable et qui servent de langue véhiculaire où il existe un besoin de compréhension entre des populations différentes (Chaudenson 1979).

Les créoles français comportent le plus grand nombre de locuteurs (10 millions selon Hazaël-Massieux 1999), suivi par le groupe des créoles anglais qui compteraient 5 millions de locuteurs (Leclerc 2001).

Les créoles français se divisent comme suit, selon les régions du monde :

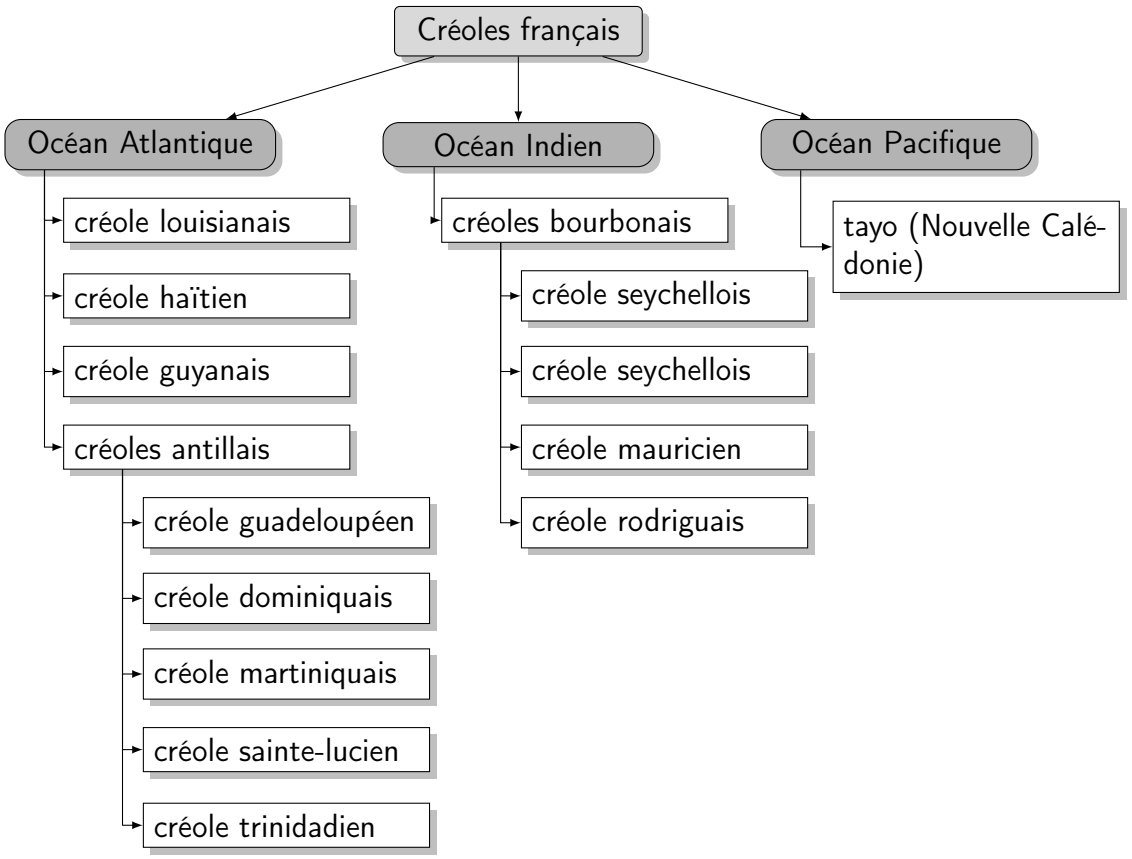


Figure 2.2 – Les créoles français selon les régions du monde.

Ci-dessous un tableau rendant compte du nombre de locuteurs de ces créoles :

Région	Nombre approximatif de locuteurs du créole
Haïti	7 000 000
Maurice	1 000 000
Réunion	600 000
Guadeloupe	425 000
Martinique	380 000
Seychelles	70 000
Rodrigues	37 000

Table 2.1 – Nombre de locuteurs des créoles (selon Leclerc 2001)

2.1.1 Présupposés existants

Les créoles ont souvent été caractérisés commes des langues « simples » (McWhorter 2001), ayant un lexique et une morphologie assez pauvres (donc longtemps considérés comme des langues inflexionnelles), et ayant des implications sociologiques qui « handicaperaient » (3) de surcroît les créolophones.

Nous ne nous pencherons pas ici sur cet aspect non-linguistique, mais nous nous efforçons de mettre en lumière qu'une multitude de travaux ont été menés pour démystifier ces *a priori*. Nous savons désormais que les créoles présentent des caractérisques plus complexes que celles qui leur étaient autrefois prêtées.

Dominique Fattier (2003) a notamment démontré des multiples phénomènes morphologiques en créole haitien (CH), qui se traduisent par l'utilisation de particules devant la forme verbale pour exprimer les distinctions de temps, d'aspect et de mode.
Michel DeGraff (2005b, 2005a) s'est également attardé à refuter les théories exceptionnalistes [2] [3] en s'appuyant sur le cas du CH.

Dans le cas du KM, la flexion verbale se traduit par deux formes [4] : une forme longue et une forme courte (Bonami et Henri 2010) – phénomène présent également dans les créoles antillais et que nous retrouvons également dans le créole louisiannais.
Mamode 2013 explique également dans un article que le mauricien présente, entre autres, une agglutination flexionelle entre le déterminant défini ou partitif et le nom, aussi bien que des phénomènes d'affixation et de réduplication dérivationnelles.

Le créole guadeloupéen présente également une flexion verbale comme susmentionné, et aussi un procédé de suffixation pour les verbes d'actions (Henri, Stump et Tribout 2017).

Enfin, il est également intéressant de noter que les locuteurs des différents créoles à base française ne se comprennent pas malgré la genèse similaire de leurs langues, ce qui peut s'expliquer par le contexte insulaire dans lequel ils existent et se sont développés. Comme l'explique Leclerc 2001, « l'intercompréhension » entre ces créoles est très limitée. Un bon nombre de facteurs tels que « l'accent, l'intonation, un nombre plus ou moins important de termes inconnus, de même que certains éléments grammaticaux et des tournures syntaxiques, peuvent entraver la compréhension, surtout lorsque les créolophones sont moins instruits. »

2. Thomason et Kaufman (1991) sont à l'origine du terme « exceptionalisme créole », une théorie qui stipule que les créoles ne sont pas de vraies langues génétiquement parlant, étant donné qu'ils se forment dans un dans un contexte de contact de langues et donc leur mode de transmission serait « abrupte » ou « anormale » (phénomène qui ne se produirait pas dans des langues à part entière.)

3. « *" Creole Exceptionalism " is defined as a set of beliefs, widespread among both linguists and non-linguists, that Creole languages form an exceptional class on phylogenetic and/or typological grounds. It also has non-linguistic (e.g., sociological) implications, such as the claim that Creole languages are a « handicap » for their speakers, which has undermined the role that Creoles should play in the education and socioeconomic development of monolingual Creolophones.* » (DeGraff 2005b.)

4. Nous détaillons la morphologie du créole mauricien à la section 2.3.4

2.2 Créoles français et TAL

2.2.1 Travaux et ressources existant·e·s

Dictionnaires électroniques et lexiques

Dans son article dans le cadre de l'atelier *Language Technology for Normalisation of Less-Resourced Languages*, Gonzalez (2012) se focalise sur la création de ressources numériques pour les créoles français sous forme d'un dictionnaire unique répertoriant les entrées de tous les dictionnaires créoles existants, couplée d'une analyse morphosyntaxique du corpus créé.

Il effectue un travail en amont en répertoriant comme suit les différentes ressources électroniques disponibles en ligne pour ces langues. Nous reprenons ici uniquement les ressources qui sont encore accessibles. De plus, nous nous focalisons ici uniquement sur les ressources dans un format facilement réutilisable : Gonzalez (2012) répertorie également les dictionnaires ou autres recueils de lexique en formats PDF par exemple, mais étant très difficilement réutilisables avec un pré-traitement peu coûteux, nous ne les listons pas dans cette section :

— **Lexilogos**[5] (161 entrées) : cette page sert de portail dirigeant vers d'autres sites axés sur différents créoles. On y trouve notamment des glossaires sur les créoles des Caraïbes (Guadeloupe, Saint-Barthélémy, Martinique, Haïti, Guyane) et de l'Océan Indien (principalement Maurice et la Réunion), qui comprennent également des exemples de variations. Il y a aussi une liste de liens vers des travaux linguistiques qui ont été effectué sur ces langues.

— **Écrit créole**[6] (90 entrées) : un lexique des créoles des Caraïbes, avec une traduction ou une définition pour chaque entrée mais aucune information morphosyntaxtique ou flexionnelle.

— **Petit lexique du créole antillais**[7] (107 entrées) : la seule information disponible pour les entrées présentes est une définition ou une traduction.

— **Antan Lontan**[8] (124 entrées) : un lexique du créole des Caraïbes par Marie-Andrée Blameble, présentant pour chaque entrée seulement une définition ou une traduction. Des exemples d'utilisation sont parfois disponible.

— **Dictionnaire créole**[9] (477 entrées) : un lexique contenant des entrées en haïtien, guadeloupéen, martiniquais et réunionnais avec leur définition. Aucune indication morphologique n'est présente, ni d'exemples d'utilisation ou de références. Il est néanmoins possible de trouver plusieurs définitions pour une même entrée.

5. http://www.lexilogos.com/creole_langue_dictionnaires.htm
6. http://ecrit.creole.free.fr/lexique.html
7. http://www.ieeff.org/creole.html
8. http://antanlontan.chez-alice.fr/motscreo.htm
9. http://www.dictionnaire-creole.com/

— **Potomitan** [10] (619 entrées) : un lexique du martiniquais par Raphaël Confiant, où chaque entrée est uniquement associé avec sa traduction française.

À cela nous pouvons ajouter le dictionnaire du créole mauricien d'Andras Rajki [11] (2100 entrées), où chaque entrée est associée avec sa traduction anglaise, des indications étymologiques et un exemple d'utilisation.

Un deuxième dictionnaire dans la même langue est également disponible sur le site du mouvement *Lalit* [12], où chaque entrée est associée à sa catégorie grammaticale et sa définition.

Il est à noter que ces deux ressources ne sont pas en créole mauricien « standard » (selon Carpooran 2011).

Outils de traduction

Dabre, Sukhoo et Bhattacharyya 2014 ont tenté de construire un traducteur automatique pour le créole mauricien se basant sur un algorithme calculant des statistiques de fréquences de succession de mots extraits de corpus multilingues (*Statistical Machine Translation* (SMT) [13]). Sachant qu'il n'existe pas encore d'outils de pré-traitement linguistique pour le créole mauricien, ils se limitent à des approches phrastiques (où ils alignent des phrases anglais → mauricien, puis mauricien → anglais ; français → mauricien, puis mauricien → français, et enfin anglais → mauricien en passant par le français).

Pour ce faire, ils ont construit manuellement des corpus parallèles mauricien-français et mauricien-anglais, étant donné que ces corpus prêts à l'emploi n'existent pas. Ils utilisent néanmoins un module SMT pré-existant.

Ils se rendent compte que la traduction anglais → mauricien en passant par le français dans une étape intermédiaire, comme une sorte de pont, présentait les résultats les plus concluants et ils l'expliquent par le fait que le créole mauricien et le français sont assez proches.

Quant à Lewis 2010, il a travaillé sur la création d'un moteur de tradution automatique pour le créole haïtien. C'est un travail qu'il a effectué avec l'équipe de Microsoft Translator. Il utilise donc leurs modèles SMT existants pour d'autres langues déjà traîtées par l'outil de traduction de Microsoft, avec quelques adaptations spécifiques au créole haïtien.

La disponibilité de corpus pré-existants pour les langues peu dotées en général étant faible, il a obtenu des corpus de deux sources principales : d'un petit corpus parallèle haïtien-anglais de l'Université Carnegie Mellon [14] et de la bible bilingue haïtien-anglais.

10. http://www.potomitan.info/dictionnaire/francais.php
11. http://web.archive.org/web/20101010120405/http://www.freeweb.hu/etymological/Morisyenweb.htm
12. https://www.lalitmauritius.org/en/dictionary.html?
13. https://en.wikipedia.org/wiki/Statistical_machine_translation
14. http://www.speech.cs.cmu.edu/haitian/

À la suite de ce projet, il réussit à intégrer le moteur de traduction haïtien à un agent conversationnel, où l'utilisateur entre des phrases en anglais et l'agent conversationnel les traduit en haïtien.

Outils de collecte de corpus annoté

Krik [15] est une plateforme de production partipative, mise en place par Alice Millour (2018), qui vise à annoter du corpus en catégories morphosytaxiques pour le créole guadeloupéen. La plateforme propose à ses utilisateurs des textes annotés par des outils de pré-annotation, et les utilisateurs valident ou non ces annotations automatiques.

Les corpus utilisés proviennent de conversations orales transcrites de la base de données COCOON [16], de proverbes de la page Wikipédia en français sur le guadeloupéen [17] de de l'incubateur Wikimédia du guadeloupéen [18].

Suite à cela, cette plateforme a été adaptée pour collecter du texte brut sous forme de recettes [19] et effectuer l'annotation de ces textes.

Les ressources créées [20] à la suite de ces projets restent assez peu conséquentes, dû au nombre assez bas de participants [21] (Millour et Karën Fort 2018).

2.2.2 Enjeux

De la partie précédente, nous nous rendons compte que les outils et ressources disponibles pour les créoles sont très peu nombreu·x·ses et d'une qualité peu satisfaisante. Les corpus existants sont incomplets, et difficilement réutilisables, et les outils – se basant sur ces mêmes corpus – atteignent une efficacité qui n'est pas encore optimale.

C'est dans ce contexte que s'inscrit notre travail, où nous espérons qu'en créant des ressources pour le créole mauricien nous pourrons contribuer plus largement à la recherche dans le TAL sur les créoles.

15. `http://krik.paris-sorbonne.fr/`
16. COllection de COrpus Oraux Numériques, voir : `https://cocoon.huma-num.fr/`.
17. `https://fr.wikipedia.org/wiki/CrÃ¨ole_guadeloupÃ©en`
18. `https://incubator.wikimedia.org/wiki/Wp/gcf`
19. `https://bisame.paris-sorbonne.fr/recettes/`
20. Le corpus annoté obtenu suite à ce travail est disponible sur `krik.paris-sorbonne.fr`.
21. Le qualité de l'annotation du corpus est liée aux nombres de participants (Millour et Karën Fort 2018).

2.3 Maurice et le *morisien*

FIGURE 2.3 – Carte en relief de l'île Maurice et de l'île Rodrigues (Source : *Perry-Castañeda Library Map Collection* [22])

2.3.1 Situation géographique

L'île Maurice est située dans l'Océan Indien, à environ 900 km à l'est de Madagascar, entre l'île de la Réunion et Rodrigues et fait partie de l'archipel des Mascareignes. Elle est l'île principale de la République de Maurice, qui comprend aussi Rodrigues, les îles Agalega et l'archipel de Saint Brandon.

Sa superficie est de 2 040 et compte environ 1 265 000 habitants (selon une estimation des Nation Unies), contre 40 000 habitants à Rodrigues (109 km^2) et 300 habitants à Agalega (24 km^2). Les îlots de Saint Brandon ne sont pas habités

22. `https://legacy.lib.utexas.edu/maps/islands_oceans_poles/mauritius_rel90.jpg`

en permanence, en dehors des fréquentes haltes de bateaux de pêche sur leur côtes (Baker et Kriegel 2013).

2.3.2 Histoire

La découverte de Maurice daterait de l'antiquité. D'après certains historiens, les Phéniciens auraient découvert l'île lors de leurs voyages lointains. (cf. Portail touristique de l'Île Maurice p.d.)

La période arabe

La conquête arabe de l'Océan Indien commença dès le XVIII^{ème} siècle, par le biais du commerce avec l'Afrique. Les navigateurs arabes établirent des comptoirs jusqu'aux rives de Madagascar et à partir de là firent la découverte de ce que nous appellons aujourd'hui les îles des Mascareignes. Ils les nommèrent avec différents noms arabes selon les époques. Maurice est retrouvée sous plusieurs noms sur leurs cartes : *Dina Robin*, *Dina Arobin*, *Dina Novare*[23]. À l'époque, l'île intéressait peu et elle n'aurait servi que de refuge aux pirates.

La période portugaise

Dès le début du XV^{ème} siècle, les Portugais sont autorisés à naviguer dans l'Océan Indien et à aborder l'Afrique. Cela va largement affecter l'emprise des Arabes sur la région. Ils s'emparent des Mascareignes et l'île Maurice est nommée *Cirné*. Là encore, ces nouveaux voyageurs trouvent l'île sans intérêt, l'utilisant seulement comme un point d'escale. (cf. Portail touristique de l'Île Maurice p.d.)

La période hollandaise

En 1598, lors d'un orage, un navire hollandais vient trouver refuge dans une baie au sud de l'île Maurice. Le commandant du navire, l'Amiral Wybrandt Van Warwyck, nommera l'île *Mauritius* en l'honneur du prince Mauritz Van Nassau.

Néanmoins, les hollandais n'occuperont l'île qu'à partir 1638 et ce pour une courte période de 20 ans. Ils auraient éprouvé d'immenses difficultés à cultiver, à être auto-suffisants en nourriture et à développer leurs habitations. Ils abandonneront donc l'île en 1710.

Ils auraient toutefois introduit la canne à sucre, des animaux domestiques et des cerfs sur l'île pendant leur occupation. (cf. Portail touristique de l'Île Maurice p.d.)

23. *Dina* viendrait du sanskrit *Dwipa* qui signifie « île ».

La période française

L'île devient une colonie française lorsque Guillaume Dufresne D'Arsel prend possession de l'île en 1715, qui était alors un parfait point d'escale sur la route vers l'Inde. Il la nommera Isle de France.

En 1721 à l'arrivée de Mahé de Labourdonnais, l'Isle de France commence à se développer : un nouveau port est construit à Port Louis, des bâtiments sont construits (certains existent toujours aujourd'hui) (Government of Mauritius p.d.).

Les esclaves sont achetés en Afrique de l'ouest, à Madagascar et en Inde, et sont amenés sur l'île pour ces travaux de construction. Dès 1730, le nombre de non-Européens dépassera considérablement le nombre d'Européens vivant sur l'île et à cette période, les langues non-européennes parlées à Maurice sont le bengali [24], l'indo-portugais, le malgache, le mandinika [25], le tamil, le wolof et le yoruba [26] entre autres (Baker et Kriegel 2013).

1715	Prise de possession de l'île (inhabitée) par la France.
1721	Début de la colonisation. Le français dialectal de l'Ouest devient la langue véhiculaire.
1734	1 450 esclaves, principalement originaires d'Afrique de l'Ouest et de Madagascar ; moins de 100 européens.
1767	Environ 20 000 habitants dont 80 % d'esclaves, principalement venus de la côte est de l'Afrique et de Madagascar (Allen 2008).
1773	Première mention textuelle d'un parler créole distinct du français, et langue usuelle des esclaves (Baker 2007).
1810	Conquête de l'île par les Anglais.

Table 2.2 – La période française (Henri et Bonami 2016)

Durant les guerres napoléoniennes, l'Isle de France servait de base d'où les corsaires français menaient des embuscades sur des bateaux de la flotte britannique jusqu'en 1810 où les Anglais ont pris l'île après une « vaste expédition » (Government of Mauritius p.d.).

24. « Le bengali ou bangla est une langue indo-iranienne de la famille des langues indo-européenes. » (voir `https://fr.wikipedia.org/wiki/Bengali`)

25. « Le mandinka est une langue mandingue et une variante du mandingue parlée en Guinée au Sénégal, en Gambie et en Guinée-Bissau. » (voir `https://fr.wikipedia.org/wiki/Mandinka`)

26. « Le yoruba (autonyme : *yorùbá*) est une langue d'Afrique de l'Ouest appartenant au groupe des langues yoruboïdes, groupe qui se rattache lui-même à la famille des langues nigéro-congolaises. » (voir `https://fr.wikipedia.org/wiki/Yoruba_(langue)`)

La période anglaise

Après la conquête de l'île par les Anglais, l'île est renommée *Mauritius*. La culture et la langue française seront sauvegardées selon les termes du transfert de propriété, et de ce fait la majeure partie des habitants français resteront.

L'administration anglaise apportera de grands changements socio-économiques, dont le plus notable étant l'abolition de l'esclavage en 1835.

Suite à cela, des travailleurs indiens seront recrutés pour travailler dans l'industrie du sucre qui était alors en expansion. Ces immigrants indiens ajoutent aux langues parlées à Maurice le gujarati, le marathi, le télougou et le bhodjpouri (Baker et Kriegel 2013).

À partir du XIX$^{\text{ème}}$ siècle, des immigrants chinois commenceront également à s'installer sur l'île (Baker et Kriegel 2013).

Dans les années 1940, un mouvement pour l'indépendance commence à émerger afin que l'île revienne aux mauriciens et soit dirigée par eux. Parallèlement, la politique de la Grande Bretagne dans les années soixante à pour but se séparer de Maurice, et de fait de toutes ses colonies. Après une période de tension, l'île obtiendra donc son indépendance et rejoindra le Commonwealth le 12 mars 1968.

2.3.3 Contexte démolinguistique actuel

La diversité linguistique à Maurice est assez impréssionante : les recensements démographiques de 2011 montrent que 87,8 % de la population parle le créole mauricien à la maison, 5,9 % le bhodjpouri, 4,7 % le français et 0,8 % l'hindi (Government of Mauritius 2011).

Langue	% de la population	Nombre
Créole mauricien	87,8	1 086 093
Bhodjpouri	5,9	72 638
Français	4,7	58 569
Hindi	0,8	9 721
Non-déclaré	0,1	1 549

Table 2.3 – Langues habituellement parlées à la maison selon le recensement de 2011

À Maurice, certaines langues sont utilisées uniquement dans un contexte religieux, tel que le sanskrit, l'arabe et le latin. Le tamoul, le télougou et le marathi ont un statut de langues ancestrales mais ne sont que très rarement parlés.

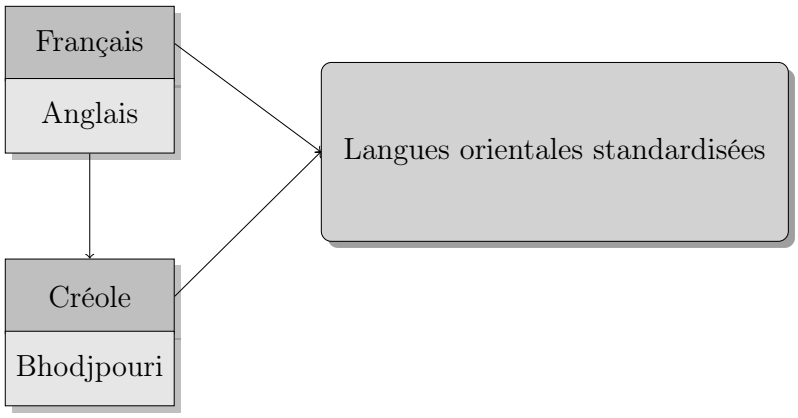
Le français est la langue la plus représentée dans les médias, et l'anglais a le statut de langue officielle depuis 1814 ; c'est la langue utilisée en politique (avec le créole), dans l'administration, dans les textes de lois, dans l'éducation. Néanmoins, l'anglais n'a pas d'usage hors de ces contextes formels et légaux. (Eriksen 1990)

Le mandarin reste présent, les sino-mauriciens (qui forment 3 % de la population) perpétuent son apprentissage mais les langues vernaculaires originelles, le hakka et le cantonais, ne sont que très rarement parlées aujourd'hui.

De ce fait, selon Eriksen 2018, seuls le créole mauricien et le bhodjpouri seraient *de facto* les langues nationales de Maurice.

C'est donc dans cet environnement de diglossie [27] [28] que se présente le créole mauricien, où il coexiste avec le français, l'anglais et de diverses langues indiennes.

On parle plus précisemment de « diglossie emboîtée » [29] :



(Baggioni et Robillard 1990)

Figure 2.4 – Schéma de « diglossies emboîtées » ou hiérarchie des langues à Maurice.

Bien entendu, les langues parlées à Maurice ne jouissent pas de la même importance ; elles sont utilisées de façon inégale et non-interchangeable, dépendant du contexte d'énonciation et du locuteur avec lequel l'échange à lieu [30]. Donc, on peut dire que ces langues sont fonctionnellement complémentaires.

27. *« Diglossia is a relatively stable language situation in which, in addition to the primary dialects of the language (which may include a standard or regional standards), there is a very divergent, highly codified (often grammatically more complex) superposed variety, the vehicle of a large and respected body of written literature, either of an earlier period or in another speech community, which is learned largely by formal education and is used for most written and formal spoken purposes but is not used by any sector of the community for ordinary conversation. »* (Ferguson 1959)

28. De fait, beaucoup de créoles existent dans un contexte digglossie, où sont parlé un créole donné et une deuxième langue. Maurice serait en ce sens un contexte de « polyglossie ».

29. voir « Les particularités du français régional de l'Île Maurice » (Desha et Wong) : `http://andre.thibault.pagesperso-orange.fr/IleMaurice.pdf`

30. De ce phénomène découle également le « code-switching ».

2.3.4 Le créole mauricien

Le créole mauricien ou *kreol morisien* (KM) s'est développé pendant la période française lors des déplacements forcés des esclaves africains et de leurs interactions entre eux aussi bien qu'avec leurs maîtres. Dès la fin du XVIII$^{\text{ème}}$ siècle, le *kreol* était la principale *lingua franca* et la langue maternelle des esclaves nés à Maurice (Baker et Syea 1991).

Après l'abolition de l'esclavage en 1839, les travailleurs indiens ainsi que les immigrants chinois venus s'installer sur l'île commenceront très vite leur apprentissage du *kreol* (dès le début du XX$^{\text{ème}}$ siècle).

Durant les années suivant l'indépendance, bien qu'omniprésent et considéré comme le parler national par la majorité de la population, un « consentement commun » (Goodchild 2013) pesait sur le KM, qui voudrait qu'il ne soit pas assez sophistiqué et développé pour être utilisé lors de communications formelles. Le KM n'avait alors pas un statut de langue à part entière ; il était considéré comme un dialecte parlé, qui n'avait pas lieu d'être écrit.

Néanmoins, depuis deux décénies environ se produit un basculement du statut de cette langue pour les Mauriciens. Il est possible d'observer une utilisation plus accrue du KM dans des contextes formels : à la télévision, dans les journaux et dans l'éducation.

En 2004, pour pousser la population à utiliser la langue à l'écrit comme à l'oral, le ministère de l'éducation et de la recherche commande la *Graphi-larmoni*[31] à Vinesh Hookoomsing (professeur de linguistique à l'Université de Maurice) pour accompagner ce changement. (Hookoomsing 2004)

Suite à cela, le gouvernement Maurcien mettra en place l'*Akademi Kreol Morisien* (AKM), un « comité technique de haut niveau, qui a pour mission d'examiner tous les aspects relatifs à l'introduction du créole mauricien comme matière optionnelle et d'élaborer, entre autres, une version harmonisée de la langue écrite. » (Harmon 2011)

Ils publieront deux rapports en 2011 : *Lortograf Kreol Morisien* et *Gramer Kreol Morisien* – entièrement rédigés en KM.

Les autorités et les associations poussent pour la diffusion et l'utilisation massive du KM dans tous les contextes socioculturels, mais une grande partie de la population peine encore à avoir accès à l'apprentissage de la graphie et de l'orthographe préconisées. Il est vrai que le KM a été introduit dans les écoles depuis 2011 et les étudiants après cette date sont donc sensibilisés à l'écriture du KM mais peu de structures sont mises en place pour l'apprentissage du KM hors du système scolaire. De ce fait, nous nous retrouvons dans une situation transitoire au niveau du KM où une partie des écrits sont rédigés suivant la graphie proposée par ceux qui ont appris à l'écrire et d'autres non.

C'est pour cela que dans ce projet, nous traiterons aussi bien les occurences en KM « standard » que les occurences présentant une orthographe libre / non-conforme au standard.

31. *Graphi-larmoni* signifie littéralement « la graphie de l'harmonie » puisqu'elle avait pour but d'harmoniser le système d'écriture du KM.

Il est intéressant de noter qu'en poussant les locuteurs à s'exprimer dans la graphie préconisée, une partie de la population pourrait se sentir exclue. Certains locuteurs semblent avoir une envie réelle de participer à ce mouvement de diffusion du KM mais se sentiraient freinés par leur manque de connaissance de la graphie préconisée (selon les résultats du sondage détaillé dans la section 3). Ils préfèrent ne pas écrire du tout que d'écrire et de se « tromper », même si cette démarche de normalisation du KM par le gouvernement ne s'inscrit pas dans un « système clos » mais a plutôt pour objectif « de livrer des propositions pour qu'elles soient testées » (Hazaël-Massieux 2005).

Classification

Le KM est un créole à base française proche du créole seychellois, du créole rodriguais et du créole chagossien.

Les avis divergent sur la relation du KM avec ces autres créoles : pour Chaudenson 1979, le KM découlerait (directement ou indirectement) d'une première forme de créole réunionnais mais pour Baker et Corne 1982, l'influence du créole réunionnais sur le KM est négligeable.

Pour défendre son postulat, Chaudenson 2013 avance que les deux créoles, bien que proches au début du XIX$^{\text{ème}}$ siècle, se sont différenciés dans la suite. Le caractère insulaire dans lequel ils se sont développés explique « leurs évolutions respectives propres » et leurs caractéristiques lexicales bien distinctes.

Comme l'explique encore Chaudenson 2013, contrairement au mauricien qui se serait « basilectalisé »[32], le réunionnais aurait subit une « érosion basilectale ». De ce fait, le réunionnais présenterait une « réduction voire la disparition de traits basilectaux, initialement communs, comme, par exemple, des agglutinations de l'article (type *lakaz*), l'imparfait en « *té* + verbe », la négation *napa*, l'emploi de *ansam* au sens de « et », *etc* ».

Français	Créole mauricien	Créole reunionnais
Peuples créoles du monde entier, donnons-nous la main.	*Tou dimoune ki koz langaz kreol anou mars ansam.*	*Anou pèp kréol dan lo Monn antyé anon mèt ansanm.*
Le créole est la puissante langue de notre patrie, car il est parlé par tout le monde.	*Langaz kreol pli gran patrimwann nou pei parski tou dimounn koz li.*	*Lo kréol lé la lang lo pli gabyé nout nasyon parské tout domoun i koz ali.*

Table 2.4 – Exemples comparatifs du créole mauricien (île Maurice) et du créole réunionnais (île de La Réunion) cf. Leclerc p.d.

32. « Le basilecte » est la variété d'une langue la plus éloignée de sa variété de prestige, l'acrolecte ». Ce terme à été proposé par William Alexander Stewart en 1965. Voir : `https://fr.wikipedia.org/wiki/Acrolecte_et_basilecte`

Caractéristiques du KM

Morphologie

Une grande majorité du vocabulaire du KM a été adapté du français et l'ordre des mots dans les phrases suit l'ordre SVO. Néanmoins, il n'a pas hérité de son système flexionnel : le KM ne montre aucune trace de flexion en temps, mode ou aspect, ni en nombre et en genre (Bonami et Henri 2010).

Ci-dessous un exemple pour illustrer ce phénomène :

(1) Mo/to/li/nou/zot manz kari.
1SG/2SG/3SG/1PL/2PL manger.FC curry
'Je/tu/il/nous/ils mange·s·ons·ont du curry.'
(Bonami et Henri 2010)

Il existe toutefois deux formes verbales en KM : la forme courte (FC) et la forme longue (FL) qui est généralement marquée par un morphème final (*i* ou *e*) (Grant 2009).

FL	bʁize	bʁije	bʁije	vɑ̃de	amɑ̃de	kɔ̃siste	ʁeste	fini	vini
FC	bʁiz	bʁij	bʁije	van	amɑ̃d	kɔ̃siste	ʁes	fini	vin
TRAD.	casser	briller	mélanger	vendre	amender	consister	rester	finir	venir

Figure 2.5 – La conjugaison à deux formes du KM (tableau de Henri et Bonami 2016).

De plus, les indications de TMA sont exprimés grâce à des morphèmes libres.

(2) Mo ti manze.
1SG TMA.PASSÉ manger.FL
'J'ai mangé du curry.'

(3) Mo pe vinn guett twa.
1SG TMA.PRESENT venir.FC voir.FC 2SG
'Je viens te voir.'

(4) Mo finn blie.
1SG TMA.PASSÉ oublier.FL
'J'ai oublié.'

(5) Mo ava dir li.
1SG TMA.FUTUR dire.FC 3SG
'Je lui dirai.'

Phonologie

La phonologie du KM est essentiellement identique à celle du français standard à l'exception de quelques phonèmes. Le /ʃ/ et le /ʒ/ du français se sont dépalatalisés en /s/ and /z/ en KM, et les voyelles antérieures /y/ and /ø/ ont perdu leur trait arrondi devenant /i/ et /e/ respectivement (Baker 1972).

Lorsque le /ʁ/ du français est précédé d'une voyelle, il se vocalise en une voyelle rhotique.

French↦Mauritian	example	trans.
ʃ↦s	detaʃe ↦detase	'detach'
ʒ↦z	mɑ̃ʒe ↦mɑ̃ze	'eat'
ʁ↦ɚ/__[$_\sigma$	paʁti ↦paɚti	'leave'
y↦i	fyme ↦fime	'smoke'
ə↦e/#C__	ʁədone↦ʁedone	'give again'
ɛ↦e	fɛʁ ↦feɚ	'do'
ɔ↦o	sɔʁti ↦soɚti	'go out'

Figure 2.6 – Les adaptations du système phonétique du KM à partir du français (Bonami et Henri 2010)

Lexique

Avec les différentes vagues de colonisation et d'immigration qu'a connu l'île Maurice, le vocabulaire du KM comprend des mots d'une variété d'origines.

Comme susmentionné, beaucoup de mots proviennent du français, mais ils ne s'utilisent pas toujours dans le même contexte et ont parfois un sens légèrement différent. Par exemple, *gagn* (qui veut dire « avoir », « obtenir » en KM) est dérivé de « gagner » et n'a donc pas le même sens.

La deuxième plus grande influence sur le lexique du KM comprend les langues indo-européennes.

Certains mots, même s'ils sont présents dans un pourcentage minime, trouvent également leurs origines dans le portugais, l'espagnole, le malgache, le tamoul et le chinois entre autres.

Selon un des premiers dictionnaires du KM de Baker et Hookoomsing 1987, 10 % du lexique du KM proviendraient de langues non-francophones.

Il est également possible de retrouver des emprunts fréquents à l'anglais et ces mots gardent en général leur orthographe anglaise et doivent être écrits entre guillemets. [33]

33. Selon les recommendations de l'*Akademi Kreol Morisien* dans le rapport sur l'orthographe, p.27 section 4.1.

Graphie

Contrairement à ses homologues antillais et réunionnais, le KM n'utilise aucun signe diacritique mais, comme tous les créoles à base française, il s'écrit avec l'alphabet latin. Ci-dessous quelques conventions orthographiques préconisées par l'AKM dans *Lortograf Kreol Mauricien* (2011) :

— Consonnes complexes
  - « gn » est la forme normale choisie, mais peut aussi retrouvé dans des textes comme « yn ». ex. *gagn* ou *gayn* (obtenir, avoir)

— Distinctions
  - « dj/j » Le choix dépend de la langue source duquel le mot provient. Quand un mot est emprunté du français, di- s'impose naturellement (ex. *media, diab, diaman, etc.*). Par contre, quand un mot provient d'autres langues (comme l'anglais ou les langues indiennes), « j » semble plus juste. (ex. *baja, jam, jak, joukal, jous, maja, etc*).
  - « i/y » pour [j] Conserve le principe du *Grafi larmoni*[34], qui dit qu'entre une consonne et une voyelle, « i » est utilisé (ex. *morisien*) et en début ou fin du mot « y » est utilisé (ex. *yer, may*)

— Traits d'union
  - nom + article défini ex. *tifi-la* (cette fille là)
  - noms composés ex. *akt-dese, aryer-granper, bien-et*

— Apostrophes
  - utilisés dans des formes élidées ex. ***mo'nn*** *koz ar li,* ***li'a*** *koz ar twa, nou* ***ti'a*** *kontan zwenn* twa.[35]

Malgré les efforts du gouvernement Mauricien pour harmoniser l'écriture du KM, cette graphie ne s'est pas encore imposée. Effectivement, dire que le KM est aujourd'hui « standardisé, » revient à assumer que les locuteurs de la langue écrivent tous le KM suivant les mêmes règles orthographiques.

Notre expérience en tant que locutrice de la langue montre bien que ce raisonnement est erroné, n'ayant jamais appris à lire et à écrire selon ces règles.

De ce fait, un des objectif de ce travail est de refléter telle qu'elle est aujourd'hui. Nous ne standardisons pas les corpus collectés et cela à aucune étape du traitement.

34. Rapport commandité par le gouvernement mauricien en 2004, Hookoomsing 2004
35. *Lortograph Kreol Morisien,* Carpooran 2011, p.43

Chapitre 3

Étude préliminaire : mieux comprendre les locuteurs

3.1 Une enquête sous forme de questionnaire

3.1.1 Format des questions

Comme un premier contact avec les locuteurs du KM, nous avons décidé de mener un sondage pour connaître la relation que les Mauriciens entretiennent avec leur langue. Nous l'avons intitulé « Le créole mauricien et sa présence en ligne » [1].

Pour créer le formulaire, nous avons utilisé Framapad [2], une instance du logiciel libre Etherpad [3] sous licence Apache 2.0 [4].

Il a été largement inspiré de l'étude menée par Alice Millour sur l'alsacien (Millour 2019) et prend une structure similaire, mais divisée en 3 parties :

1. Le créole mauricien et vous
2. Le créole mauricien, internet et vous
3. Participer au développement du créole mauricien grâce à vos connaissances

La première partie nous renseigne sur le profil des participants et leur auto-évaluation sur leur niveau de KM, la deuxième sur ce qu'ils écrivent en KM sur Internet, et la troisième sur leur opinion sur la production participative de ressources.

Nous avons également fait le choix de ne pas intégrer une option « je ne sais pas/neutre » autant que possible afin de forcer nos participants à la réflexion et ne pas choisir cette option par défaut, car cela ne nous donnerait pas d'information utile pour nos analyses.

Toutes les réponses collectées sont strictement anonymes.

1. Le formulaire est accessible à :
`https://framaforms.org/sondage-le-creole-mauricien-et-sa-presence-en-ligne-1555054850`
2. Voir : `https://framapad.org/fr/`
3. Voir : `https://etherpad.org/`
4. Voir : `www.apache.org/licenses/LICENSE-2.0.html`

3.1.2 Objectifs du sondage

Ce travail s'inscrivant dans la suite des travaux d'Alice Millour sur l'alsacien et le créole guadeloupéen (Millour et Karën Fort 2018), un des objectif du sondage était de comprendre l'échec relatif de la campagne sur le guadeloupéen, pour laquelle il n'y a pas eu d'enquête sous forme de sondage au préalable (sur la plateforme dédiée au créole guadeloupéen, 1 205 annotations ont été produites par seulement 11 participants dans une période de 9 jours).

Le sondage sur l'alsacien avait été conduit après la mise en place de la plateforme et après l'analyse de résultats obtenus. Il avait connu un franc succès, avec 1 224 réponses en deux mois (2019), mais les retours de l'enquête n'ont pas pu être pris en considération lors de la mise en place de la plateforme (2016).

De ce fait, en amont de la mise en place de la plateforme de production participative pour le KM, il nous a semblé important de connaître les habitudes des mauriciens par rapport à l'usage de leur langue sur Internet et jauger leur disposition à participer à la création de ressources pour le KM. Ce sondage a été donc conduit avec l'intention de :

— connaître le contexte dans lequel les mauriciens utilisent le KM en ligne (s'ils l'utilisent, pour quoi faire, *etc.*),
— observer leur relation par rapport à leur langue, savoir s'ils sont au courant des tentatives de normalisation du KM et s'ils se tiennent aux prescriptions de l'*AKM* sur l'orthographe et des ressources existantes en KM en général,
— avoir un premier contact avec des locuteurs du KM et les sensibiliser à la production participative de ressources,
— déterminer comment adapter la plateforme selon les attentes de participants.

Nous espérions aussi que faire la promotion de la plateforme auprès des répondants du sondage augmenterait la participation.

Une première diffusion a été faite par e-mail le 15 avril 2019, puis sur les réseaux sociaux (*via* une publication sur notre page personnelle Facebook), puis l'envoi d'un message privé aux contacts Facebook étant locuteurs du KM le 23 avril 2019. Au 31 juillet 2019, 143 personnes avaient répondu au sondage.

3.2 Analyse des résultats

3.2.1 « Le créole mauricien et vous »

Dans cette section, nous remarquons que 74,1 % des participants sont des femmes et que la moitié des participants ont moins de 30 ans. Nous pouvons observer la distribution d'âge des participants dans le tableau suivant :

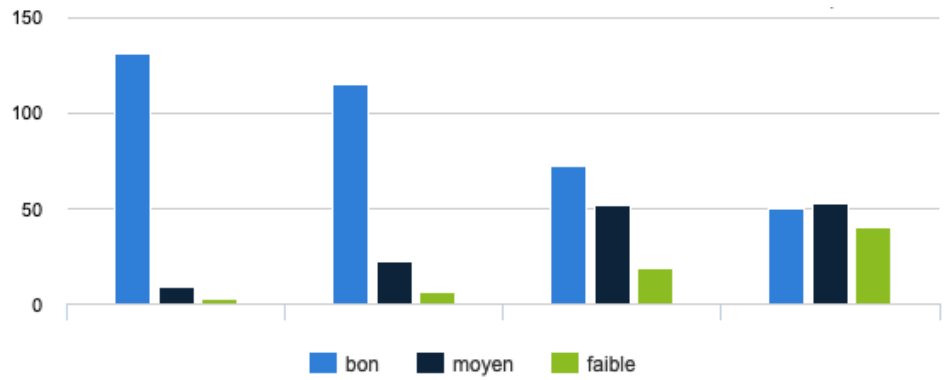
< 20	20 à 29	30 à 39	40 à 49	50 à 59	60 à 69	70 à 79	ND
15	56	33	15	19	3	1	1

TABLE 3.1 – Distribution de l'âge des participants

Parmi eux, moins de 10 % des participants ne considèrent pas le KM comme une de leur langue maternelle. 56 % d'entre eux habitent dans le centre de l'île [5], où se situes les principales agglomérations.

La majorité des personnes interrogées (56 %) n'ont pas d'investissement professionnel et/ou associatif lié au KM.

Lors de l'auto-évaluation des participants de leur niveau en KM, seulement 35 % jugent leur niveau de production écrite comme « bon » et 60 % d'entres voudraient donc l'améliorer.



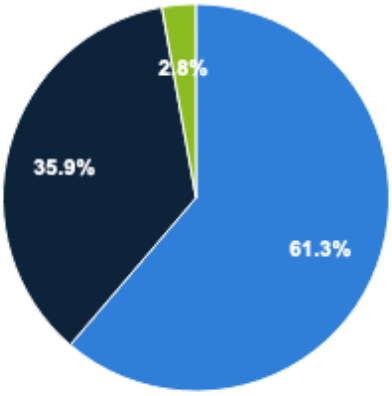
	bon	moyen
Compréhension orale	131	9
Production orale	115	22
Compréhension écrite	72	52
Production écrite	50	53

FIGURE 3.1 – Auto-évaluation des participants

5. Afin de pouvoir donner une analyse fiable sur d'éventuelles variations existantes selon les régions de l'île, il nous a semblé important d'avoir une idée d'où resident les participants.

À la question « Pensez-vous que le créole mauricien possède des variantes ? », 76 % des participants ont répondu « Oui, le créole parlé varie selon les régions de l'île ». Avant de se lancer dans ce sondage, nous avions émis l'hypothèse que le KM parlé était le même sur toute l'île et nous avons été surpris de la réponse obtenue.

Les résultats de la question suivante nous donnent des pistes pour d'autant plus explorer le phénomène de variation existant dans le KM. Nous demandons aux locuteurs de déterminer quelle orthographe leur est la plus facilement déchiffrable à la lecture d'une phrase simple (« Je suis allé·e acheter du pain »). La première option est plus proche de la forme standard, et la deuxième plus proche de la graphie française.



monn al aste dipin	87
monne alle acheté dipain	51
je ne suis à l'aise avec aucune de ces orthographes	4

FIGURE 3.2 – Réponses à la question « Laquelle de ces phrases vous semble la plus facile à lire ? »

Nous voulions en premier lieu voir avec cette question si les locuteurs du KM qui n'étaient pas familiers avec la graphie préconisée auraient plus tendance à écrire le KM avec la graphie française lorsqu'ils ne savent pas écrire un mot donné. Ceux qui ne sont pas à l'aise avec les orthographes proposées ont fourni les orthographes alternatives suivantes :

1. *mo'nn al aste dipin*,
2. *mone al aster dipain*,
3. *mone ale acheté du pain*.

La première proposition alternative est orthographiée selon les recommandations de l'AKM, avec l'apostrophe pour marquer l'élision possible entre le pronom *mo* et la particule TMA *finn*, qui se transforment en la forme contractée *mo'nn* (« j'ai »).

La deuxième, qui tend vers la graphie française, montre que son énonciateur considère le « e » final de *mone* (« j'ai ») comme muet [6] (comme souvent en français). Cela explique aussi la présence du « r » en fin de *aster* [aste] (« acheté »). Toutefois en KM, *aster* orthographié ainsi veut dire « maintenant » (prononcé [aster]) mais le contexte aide à la bonne compréhension du verbe malgré l'ambiguïté possible.

La troisième phrase présente la même tendance vers l'orthographe française avec des « e » muets en fin des deux premiers mots de la phrase. Le reste de la phrase est écrite entièrement avec la graphie française. Nous pouvons supposer que ce locuteur a opté pour la graphie française car il·elle n'était pas sûr·e de se faire comprendre.

Malgré ces différentes graphies, nous pouvons définir assez assurément que cette variation n'est pas régionale (les deux premières alternatives ont été proposées par des participants habitants le centre de l'île, et la dernière de l'ouest), elle est purement une variation orthographique, qui représente bien la situation du KM aujourd'hui – où plusieurs graphies co-existent.

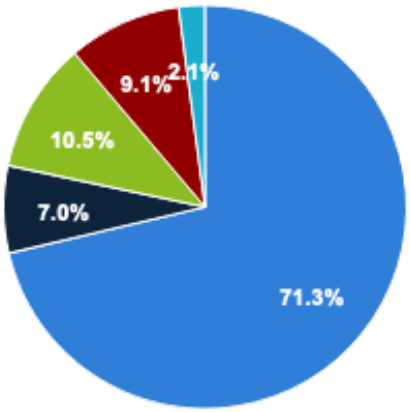
6. En KM, le « e » en finale exprime un [e].

3.2.2 « Le créole mauricien, Internet et vous »

Dans cette deuxième partie, nous en apprenons plus sur la manière dont les mauriciens utilisent le KM sur internet et hors ligne également.

67,8 % des participants affirment utiliser le KM sur Internet, même rarement (lecture et écriture compris).

L'utilisation du KM à l'écrit semble assez répandue auprès des participants : 71,3 % d'entre eux écrivent en KM (en ligne ou non).



Oui (même rarement)	102
Non, car le créole est une langue orale que je ne souhaite pas écrire	10
Non, car je ne saurais pas comment l'écrire	15
Non, car je n'en ai pas l'occasion	13
Non, pour une autre raison	3

FIGURE 3.3 – Réponses à la question « Écrivez-vous le créole mauricien (en ligne ou non) ? »

Il y aurait donc un réel besoin pour la majeure partie des Mauriciens de s'exprimer dans leur langue à l'écrit.

Les conversations sur les réseaux sociaux sont le contexte dans lequel une grande partie des participants (63,6 %) écrivent en KM.

Étant au fait que les participants pourraient éprouver des difficultés à orthographier certaines occurrences en KM, nous leur avons demandé comment ils écrivaient un mot dont ils ne sont pas sûrs de l'orthographe. 50 % d'entre eux l'écrivent comme ils l'entendent, et 30 % l'écrivent avec la graphie française.

Pour les 20 % restants, certains vérifient l'orthographe dans le dictionnaire du KM, d'autres l'écrivent tantôt comme ils l'entendent, tantôt avec la graphie française et cela dépendrait de facteurs aussi nombreux que le nombres de locuteurs.

Les facteurs qui influencent leurs choix selon certains participants sont :

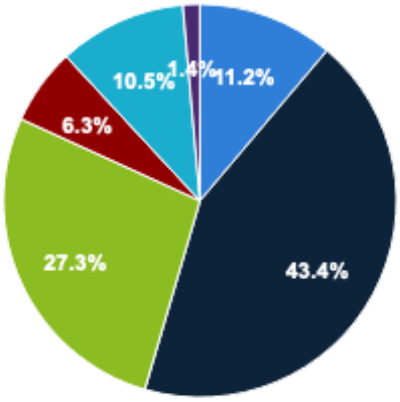
— « Si le même mot écrit en français transmet la même intention, il sera utilisé. Dans l'autre cas, ce sera comme je l'entends. »
— « Si le mot est couramment utilisé en créole, je ferai l'effort d'écrire. Si c'est un mot plutôt français, non. (ex : cancer du sein). »
— « La sonorité et comment je peux intégrer (le mot français) en créole. »
— « Je vérifie dans le dictionnaire ou selon mes connaissance en morphosyntaxe. »

Ces réponses illustrent le fait que 27 % des participants n'avaient pas connaissances de l'existence du rapport *Lortograf Kreol Morisien* (Carpooran 2011).

3.2.3 « Participer au développement du créole mauricien grâce à vos connaissances »

L'enjeu dans cette partie était de donner un aperçu de la production participative aux participants et de déterminer leurs attentes s'ils contribuaient à un tel projet.

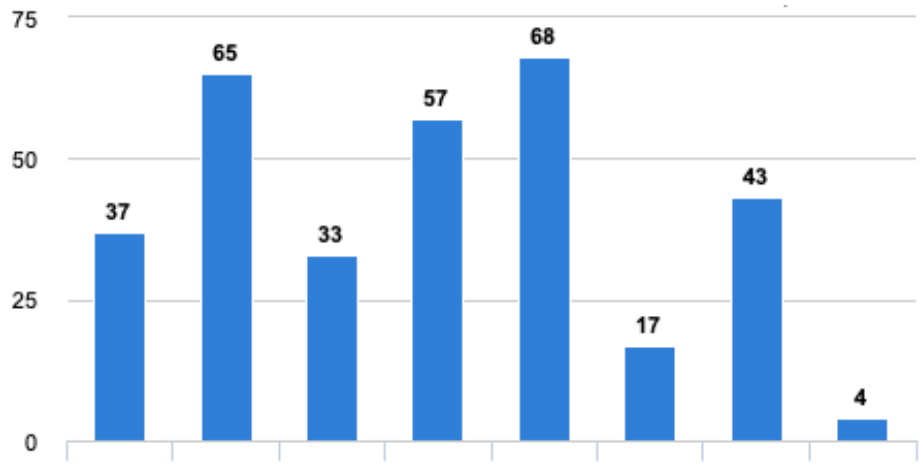
Bien que 72,7 % ne connaissaient pas le principe, 81,9 % d'entre eux trouvent que c'est une bonne idée de créer des ressources pour le KM par production participative.



Une bonne idée, je le fais déjà !	16
Une bonne idée, mais je ne sais pas comment faire.	62
Une bonne idée, mais je n'ai pas le temps.	39
Trop compliqué pour moi, je ne suis pas à l'aise avec l'informatique	9
Trop compliqué pour moi, mon niveau de créole n'est pas assez bon	15
Une mauvaise idée (expliquez pourquoi ci-dessous)	2

Figure 3.4 – Distribution des réponses à la question « Participer à la production collaborative de ressources en ligne pour le créole mauricien vous paraît ».

La figure suivante résume ce que les participants attendent d'un projet de création de ressources pour le KM :



de rentrer en contact avec d'autres personnes parlant le créole mauricien pour pouvoir discuter	37
de partager des contenus, des conseils, des astuces, des avis	65
de vous amuser en ligne	33
d'améliorer votre créole	57
d'apprendre des choses (de manière générale)	68
de gagner des bons de réduction (librairies ou événements culturels par exemple)	17
de valoriser votre connaissance du créole en participant à un projet de recherche	43
autre (préciser)	4

FIGURE 3.5 – Distribution des réponses à la question « Vous aimeriez que votre participation à la création de ressources en ligne vous permette… ».

Des participants motivés ?

Avec ce sondage, nous nous rendons compte que les participants semblent adhérer à l'idée de la production participative de ressources pour le KM. Nous avons reçu des retours majoritairement positifs de participants qui saluent l'initiative, et qui aimeraient pouvoir utiliser le KM plus largement dans leur quotidien.

Néanmoins, seulement 39,2 % des participants souhaitaient être tenus au courant des résultats du sondage et 27,4 % d'entre eux souhaitaient être avertis lors de la mise en place de la plateforme de production participative pour le KM.

Grâce au sondage, nous avons donc récupéré 53 adresses e-mail vers lesquelles nous pourrons diffuser le site de production participative.

Il faut noter que le sondage sur l'alsacien (MILLOUR 2019) avait été conduit après la mise en place de Bisame, la plateforme de production participative pour l'alsacien. Le sondage avait obtenu plus de 250 réponses mais les résultats du sondage n'ont donc pas pu être pris en compte étant donné que la plateforme avait déjà été lancée.

Nous espérons donc tirer parti du sondage sur l'alsacien, et également intégrer au maximum l'apport de ce présent sondage lors de la mise en place de « *Ayo* ! », la plateforme de collecte de corpus annoté en parties du discours pour le créole mauricien.

Chapitre 4

« *Ayo* ! », une application de production participative

4.1 Méthodologie

4.1.1 La recherche de corpus

Nous voulions obtenir des textes les plus diversifiés possible, afin d'observer la langue sous plusieurs aspects pour rendre compte à la fois des habitudes langagières des locuteurs natifs et de la variation du KM. Cela permettrait également que nos outils soient capables de traiter un plus grand nombres de textes.

Conformément à nos attentes, la recherche de corpus existants en KM fut fastidieuse. Le manque d'intérêt et de financement pour la technologisation des langues peu dotées en général peuvent expliquer cela (Barbaresi 2013).

Nos options se sont d'autant plus réduites dû aux formats peu exploitables de certains textes (en PDF notamment). C'est le cas des publications du mouvement politique *Lalit*[1] publie périodiquement des revues en KM sur son site, mais également des poèmes traduits en KM. En raison des contraintes de temps, nous ne pouvions pas les convertir en un format plus adapté.

La Bible, le Coran, et le *Bhagavad-Gita*[2] sont librement disponibles intégralement en KM mais nous avons fait le choix de ne pas les utiliser, car nous voulions rester le plus neutre possible afin de ne pas perdre de participants.

Nous nous retrouvons donc avec deux sources principales, exploitables et libres de droits :

1. le contenu de l'incubateur Wikimédia du KM[3],
2. les textes (proses[4] et poèmes[5]) de Dev Virahsawmy, homme politique, dramaturge, poète et linguiste.

1. `https://www.lalitmauritius.org/`
2. Ce texte est un des écrits fondamentaux de l'hindouisme.
3. Un incubateur Wikimédia est une plateforme où une communauté linguistique peut reprendre un projet Wikimédia (Wikipédia, Wiktionnaire, Wikilivres, Wikinews, Wikiquote et Wikivoyage) dans une nouvelle langue. Le lien vers la page dédiée au KM est accessible à : `https://incubator.wikimedia.org/wiki/Wp/mfe`
4. `https://boukiebanane.com/table-of-contents-konteni/proz-literer-dev-devs-literary-prose/`
5. `https://boukiebanane.com/table-of-contents-konteni/poezi-dev-devs-poetry/`

4.1.2 Le corpus brut

L'incubateur Wikimédia pour le KM référence actuellement 78 articles, tous faisant parti de 5 thèmes : *Moris*, *Geography*, *Languages*, *Relizion ek Krwayans*, *Universitity Topics*. La dernière modification date du 29 janvier 2017. Les articles sont pour la plupart très courts, et ne sont constitués que d'une seule phrase.

Il est intéressant de noter que sur la page d'acceuil de l'incubateur Wikimédia, il y a un avertissement qui impose aux participants d'écrire uniquement en KM standard et de remplacer les orthographes « non-standards » dans les pages existantes.

Des textes présents sur *Boukie Banane*, le site de Dev Virahsawmy, nous avons sélectionné deux textes en prose : 4 phrases (48 tokens) tirées de « *3 tizistwar pou lekran* » [6] et 60 phrases (606 tokens) de « *Fennsifer* » [7].

Nous avons fait le choix d'utiliser uniquement des écrits sous forme de prose pensant que la majorité des textes collectés sur la plateforme seraient d'un style proche de la prose.

De plus, nous avons utilisé 3 phrases d'exemples (32 tokens) du dictionnaire en ligne – *Morysien Dictionary* d'Andras Rajki [8], un conte produit par une locutrice du KM (2 phrases pour 43 tokens), un extrait de 2 phrases d'un poème de Bertolt Brecht (19 tokens) traduit en KM par Lindsey Collen [9] et 2 phrases (9 tokens) d'exemples tirées de *Lortograf Kreol Morisien* (Carpooran 2011).

Le tableau ci-dessous résume le contenu du corpus brut (C_{Brut}) :

	Genre	Nombre de tokens	Nombres de phrases
Boukie Banane	Prose	654	60
Wikimédia	Descriptions courtes	267	10
Texte libre	Conte	43	2
Morysien Dictionary	Exemples	32	3
Lalit	Poème traduit	19	2
Lortograf Kreol Morisien	Exemples	9	2
Total		**1 024**	**79**

Table 4.1 – Description du corpus brut

6. http://boukiebanane.com/3-tizistwar-pou-lekran/
7. http://boukiebanane.com/fennsifer/
8. http://web.archive.org/web/20101010120405/http://www.freeweb.hu/etymological/Morisyenweb.htm
9. https://www.lalitmauritius.org/modules/docpool/files/x-doc-lindsey-pu-bann-ki-puvinn-apre-nu.pdf

4.1.3 Choix du jeu d'étiquettes

Nous avons manuellement annoté en parties du discours le corpus brut, C_{Brut}, en format Brown, avec *Universal POS Tags* (Petrov, Das et McDonald 2011) – un jeu d'étiquettes faisant parti du projet *Universal Dependencies* (UD)[10] (McDonald et al. 2013). Ce sont les mêmes étiquettes utilisées dans le travail d'Alice Millour et Karën Fort sur le guadeloupéen (2018), que nous reprenons afin de capitaliser sur ce qui a déjà été fait. De plus, cela facilite l'adaptation d'une partie de la plateforme pour le KM et nous permettra ensuite d'avoir des points de comparaison sur l'évaluation entre les deux plateformes.

Classes ouvertes	Classes fermées	Autre
ADJ	ADP	PUNCT
ADV	AUX	SYM
INTJ	CCONJ	X
NOUN	DET	
PROPN	NUM	
VERB	PART	
	PRON	
	SCONJ	

Table 4.2 – *Universal POS tags*[11]

4.1.4 Enrichissement du jeu d'étiquettes

L'annotation du C_{Brut} a été effectuée sans tokénisation préalable. Nous voulions observer le corpus à travers l'annotation manuelle, et déterminer ensuite les prétraitements à effectuer, notamment la tokénisation[12].

Nous nous sommes rapidement rendues compte que les étiquettes existantes n'illustraient pas l'entièreté des phénomènes observés dans le corpus, notamment les occurrences construites par agglutination ou par élision. De ce fait, nous en avons ajouté 8 supplémentaires afin de rendre l'annotation plus intuitive pour les locuteurs du KM :

- PART_TMA
  ex. : *Bann profeser **ti** mont bann korperativ lekol ek kolez.*
  « Les enseignants ont mis en place des coopératives scolaires. »

10. « *Universal Dependencies (UD) is a project that is developing cross-linguistically consistent treebank annotation for many languages, with the goal of facilitating multilingual parser development, cross-lingual learning, and parsing research from a language typology perspective. The annotation scheme is based on an evolution of (universal) Stanford dependencies (de Marneffe et al., 2006, 2008, 2014), Google universal part-of-speech tags (Petrov et al., 2012), and the Interset interlingua for morphosyntactic tagsets (Zeman, 2008). The general philosophy is to provide a universal inventory of categories and guidelines to facilitate consistent annotation of similar constructions across languages, while allowing language-specific extensions when necessary.* »
Voir : `https://universaldependencies.org/introduction.html`

11. `https://universaldependencies.org/u/pos/all.html`

12. Nous détaillons cela dans la section 4.1.5

- ADV+PART_TMA
  ex. : *Personn **pa'nn** dir twa tap laport.*
  « On ne t'a pas dit de tocquer. »
- NOUN+DET
  ex. : ***Zanfan-la** pa pe ekout so mama.*
  « L'enfant n'écoute pas sa mère. »
- NUM+NOUN
  ex. : *Depi **wit-er** gramatin ziska **sink-er** tanto.*
  « De huit heures du matin à cinq heures de l'après-midi. »
- PART_TMA+PART_TMA
  ex. : ***Ti'a** bon gagn zot lopinion.*
  « Nous espérons avoir leur opinion. »
- PRON+ADV
  ex. : ***Mo'si** mo enn bouro ?*
  « Suis-je également un bourreau ? »
- PRON+VERB
  ex. : *Toutswit si **to'le**.*
  « Dès maintenant si tu le veux. »
- PRON+PART_TMA
  ex. : ***Mo'nn** konpran.*
  « J'ai compris. »

Comme souligné dans le travail sur le guadeloupéen (Millour et Karën Fort 2018), les étiquettes que nous avons ajouté reflètent les besoins rencontrés dans notre corpus de référence. Ce jeu d'étiquette n'est aucunement définitif, ne représentant sûrement pas la totalité des orthographes et variantes présentes à Maurice.

Une fois annoté, la distribution des étiquettes du corpus (C_{Annot}) se présente ainsi :

ADJ	ADP	PUNCT	ADV	AUX	SYM	INTJ	CCONJ	X
5,4 %	5,8 %	12,8%	5,8 %	0%	0,1%	0,3 %	2,5 %	0 %
NOUN	DET	PROPN	NUM	VERB	PART	PRON	SCONJ	PART TMA
21,1 %	4,6 %	2,5 %	1 %	14,7 %	0 %	10,2 %	4,3 %	7,9 %
ADV+ PART TMA	NOUN+ DET	NUM+ NOUN	PART TMA+ PART TMA	PRON+ ADV	PRON+ VERB	PRON+ PART TMA		
0,1 %	0,1 %	0,3 %	0,1 %	0,1 %	0,1 %	0,3 %		

Table 4.3 – Distribution des étiquettes du corpus annoté (C_{Annot}).

Nous n'utilisons pas les catégories AUX et PART, ayant déterminé – après observation du corpus – que les particularités du KM ne s'y prêtent pas. À leur place, nous utilisons l'étiquette PART_TMA pour les particules de temps, mode et aspect.

L'étiquette X garde la même utilisation décrite dans le guide *Universal POS Tags*[13] : elle servira à l'annotation des occurences qui ne correspondent à aucune catégorie existante. Les cas de *code-switching* ne font pas partie de cette catégorie.

Nous nous retrouvons donc avec 23 étiquettes.

13. https://universaldependencies.org/u/pos/all.html#al-u-pos/X

4.1.5 Tokénisation

S'étant familiarisé avec le corpus de référence ($C_{Annot} = C_{Ref}$) à travers l'annotation, nous avons fait les choix suivants :

- Nous ne séparons pas les formes élidées contenant une apostrophe et formes composées liées d'un trait d'union.

  Nous avons observé que certaines occurrences, lorsqu'elles sont scindées, produisent des unités qui n'ont pas de sens seules. Cela inclut notamment toutes les indications d'heures :

  (1) *enn-er* (« une heure »)
  **enn er*

  (2) *de-z-er* (« deux heures »)
  **de z er*

  Ici, *-er* n'est pas un morphème autonome (le morphème pour « une heure » en KM est *ler* et donc *enn ler* est correct.). Dans le 2ème exemple, « -z- » sert uniquement d'agent de liaison et n'est donc pas un morphème libre.

  Similairement, la forme élidée de la particule TMA *finn* (*-nn*) ne peut s'utiliser seule :

  (3) ***mo'nn*** *fini manze* (« j'ai (déjà) mangé »)
  **mo nn fini manze*

  Néanmoins, ce phénomène n'est pas généralisé : certaines occurrences pourraient produire des morphèmes libres si elles étaient découpées.

  C'est le cas des mots composés par réduplication (ex. *brit-brit* « brutalement, à la hâte »), des formes simples alternées (ex. *mars-marse* « flâner »), des formes contradictoires (ex. *ale-vini* « va et vient »).

  Nous avons tout de même décidé de ne pas les découper parce que cela ne représenterait pas toute la diversité lexicale du KM. Par exemple, *brit-brit* et *brit* sont deux adjectifs différents exprimant une échelle d'intensité différente.

- Nous n'effectuons pas de regroupement.

  Nous pensions, dans un premier temps, regrouper les locutions adverbiales comme *akoz samem* (« c'est pour cela »). Étant donnés les deux morphèmes qui la compose ont du sens en tant qu'unités autonomes, nous ne les avons finalement pas agglutinées pensant également qu'il serait plus naturel pour les participants de les annoter séparément.

  Concernant les structures NOM+ARTICLE DÉFINI, elles se présentent sous deux formes dans notre corpus : dans certains cas elles sont liées par un trait d'union et dans d'autres non :

(4) *garson-la* (« le garçon »)
garson la

Cela laisse à penser que certains locuteurs utiliseraient le trait d'union, et d'autres non. Normaliser en régroupant ce type d'occurrence ne rendrait pas compte de l'état actuel de la langue.

De plus, il est possible que des mots soient placés entre le nom et le déterminant :

(5) *sa garson intelizan la* (« le garçon intelligent ») (Carpooran 2011)

Dans cet exemple tiré de *Lortograf Kreole Morisien* (2011), le déterminant porte sur *garson* et non sur *intelizan*. Nous ne pouvons donc pas effectuer de regroupement dans ce cas (dans l'hypothèse où nous n'avions pas pris le parti de ne pas altérer la représentation de la langue).

En définitive, l'outil de tokénisation considère comme tokens tout ceux qui sont séparés par des signes de ponctuations excluant les apostrophes et les trait d'unions.

4.1.6 Rédaction du guide d'annotation

Afin d'accompagner les participants lors du processus d'annotation, nous avons rédigé un guide d'annotation pour chaque élément de notre jeu d'étiquettes.

Ce guide propose :

- une indication pour repérer la catégorie grammaticale en cours,
- des exemples d'utilisation de chaque étiquette
- des exemples identifiant les sources de difficultés et d'ambiguïtés susceptibles d'être rencontrées pendant l'annotation, signalées dans une section « ATTENTION ».

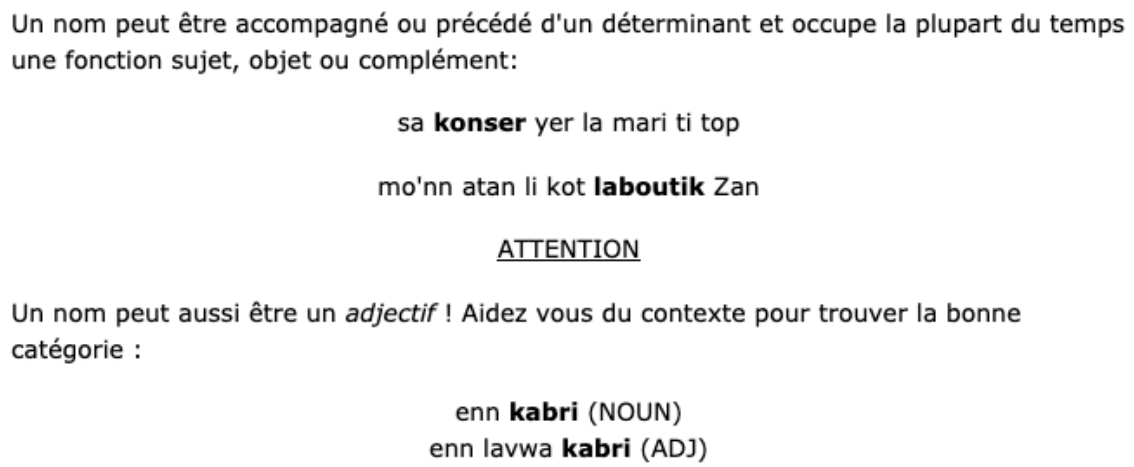


Figure 4.1 – Exemple tiré du guide d'annotation pour la catégorie noun.

4.1.7 Pré-annotation avec MElt

La pré-annotation est une méthode qui a été mise en place pour réduire le coût élevé de l'annotation manuelle par des experts. Comme le démontrent Karën Fort et Benoıt Sagot 2010, la pré-annotation augmente considérablement la vitesse de l'étiquetage morphosyntaxique.

De ce fait, tout les textes qui sont déposés par les participants sont automatiquement pré-annotés par notre outil. Les participants n'ont qu'à valider ou invalider l'étiquette proposée.

Notre outil de pré-annotation est MElt (*Maximum-Entropy Lexicon-enriched Tagger*, Denis et Benoıt Sagot 2009), un « système discriminant d'étiquetage en parties du discours » (Benoît Sagot 2016).

Pour entraîner notre modèle, nous avons divisé C_{Ref} en deux sous-corpus : le premier est composé de 849 tokens et sert de corpus d'entraînement (C_{Train}) et le deuxième, composé de 175 tokens, sert de corpus d'évaluation (C_{Test}).

Lors de l'entraînement de MElt, nous le couplons avec un lexique externe que nous avons extrait du dictionnaire *Lalit*[14].

Ainsi, nous obtenons un fichier text en format TSV (*Tab-Separated Values*) contenant 15 295 entrées lexicales associées à leur catégorie grammaticale respective. Puis, nous avons remplacé les catégories grammaticales propres au dictionnaire par celles de notre jeu d'étiquettes personnalisé pour le KM. Les entrées correspondant à plus d'une étiquette ont été dupliquées à l'aide d'un script Python.

Avec ces paramètres, MElt atteint un taux d'exactitude de 76 % sur C_{Test}. MElt est attribué un score de confiance de 76 % ce qui signifie que l'étiquette attribuée lors de la pré-annotation d'un token est correcte à 76 %.

14. `https://www.lalitmauritius.org/en/dictionary.html?`

4.2 La plateforme « *Ayo* ! »

Ayo [15] est une adaptation de Krik [16] (Millour et Karën Fort 2018) avec un double objectif :

- la collecte de corpus en KM,
- l'annotation en partie du discours de ce corpus

Contrairement aux plateformes dédiées à l'alsacien (Millour et Karën Fort 2016) et au guadeloupéen (Millour et Karën Fort 2018), les utilisateurs peuvent contribuer au projet en déposant du texte sous plusieurs formes, notamment des recettes, poèmes, proverbes. Pour ne pas limiter les productions écrites à des cases spécifiques, il est aussi possible de déposer tout autre type de texte dans un format « texte libre ».

Les utilisateurs peuvent également annoter leur propre texte ou celui des autres participants. Cette étape – la phase d'annotation – permet d'entraîner un étiqueteur morphosyntaxique et de construire ainsi un corpus annoté en parties du discours en KM.

Néanmoins, cette phase n'est pas obligatoire. Un utilisateur peut donc décider de déposer du texte et de ne pas l'annoter. S'il le décide l'utilisateur peut également ne pas déposer du texte et seulement annoter les contributions déjà faites ou améliorer les annotations existantes et proposer des variantes.

Figure 4.2 – Page d'accueil de la plateforme.

Enfin, il est également possible de proposer des variantes orthographiques depuis la page d'accueil en cliquant sur Ajouter des variantes ou en cliquant directement sur un mot dans le nuage sur cette même page.

15. « *Ayo* ! » est une interjection difficilement traduisible, qui exprime l'étonnement, la joie, la douleur entres autres. Cette expression proviendrait du tamoul *ayyo*, signifiant « hélas ». Voir : `https://en.wikipedia.org/wiki/Mauritian_Creole#Tamil_loanwords`

16. L'adaptation de la plateforme pour le KM a été effectuée par Alice Millour. Krik (plateforme pour collecter du texte annoté en guadeloupéen) est lui-même une adaptation de Bisame (alsacien), et le code source de cette plateforme est accessible ici : `https://github.com/alicemillour/Bisame`

4.2.1 L'appel à participation

La première diffusion de la plateforme a eu lieu le 9 juillet 2019 par l'intermédiaire d'une connaissance habitant sur l'île, qui a ensuite partagé le lien vers le site à son entourage.

Précédemment à cela, nous sommes allées à la rencontre des locuteurs afin de leur parler de notre travail. Durant un court séjour à Maurice, nous avons rencontré 4 personnes ayant un investissement professionnel lié au KM :

- Shrita Hassamal [17], professeure au *Mauritius Institute of Education* (MIE) et formatrice des enseignants du KM,
- Muhsina Alleesaib, professeure de linguistique et de KM à l'Université de Maurice,
- Jimmy Harmon [18], chercheur indépendant dans le domaine des langues, des cultures et des identités,
- Arnaud Carpooran, doyen de l'UFR de sciences sociales et humaines de l'Université de Maurice, auteur du *Diksioner morisien* (le premier dictionnaire du créole mauricien), des rapports *Gramer Kreol Morisien* (2011) et *Lortograf Kreol Morisien* (2011).

Ravis de partager un même intérêt pour le KM, ils ont accepté de diffuser le lien vers la plateforme à leur réseau de contacts. Suite à ces rencontres durant le mois de juin, nous les avons contacté par mail en guise de relance le 23 juillet 2019.

Pour la diffustion sur les réseaux sociaux, nous avons rédigé une publication et une affiche [19]. Nous les avons partagés sur notre page personnelle Facebook le 25 juillet 2019, et sur des groupes Facebook constitués de locuteurs du KM le 2 août 2019.

Nous avons également effectué une diffusion le 7 août 2019 auprès des 53 répondants au sondage qui souhaitaient faire partie du projet. Enfin nous avons contacté plusieurs bloggeurs et influenceurs sur Instagram, mais nos messages sont restés sans réponses.

4.2.2 La phase de formation

Bien qu'ayant constitué un corpus pour la formation, l'intégration de cette fonctionnalité n'a pas pu être effectuée à cause de complications techniques.

La phase de formation n'aurait pas été obligatoire pour toutes les étiquettes mais uniquement celles qui présentent une ambiguïté quelconque. Par exemple, peu de difficultés existent à annoter les noms propres, les nombres et les symboles mais reconnaître les adjectifs des noms communs nécessite l'aide du contexte.

Les participants peuvent néanmoins se référer au guide de formation à tout moment pendant le processus d'annotation.

17. `http://www.llf.cnrs.fr/fr/Gens/Hassamal`
18. `https://independent.academia.edu/JimmyHarmon`
19. Voir Annexe A.1.

4.2.3 La phase d'annotation

Dans cette phase, l'utilisateur corrige les annotations effectuées par notre outil de pré-annotation. Le texte qui est en train d'être annoté est affiché dans son intégralité, et il n'est possible que de corriger des mots portant une même étiquette à la fois. Pour passer aux catégories suivantes, il est obligatoire de valider ou invalider toutes les étiquettes de la catégorie en cours.

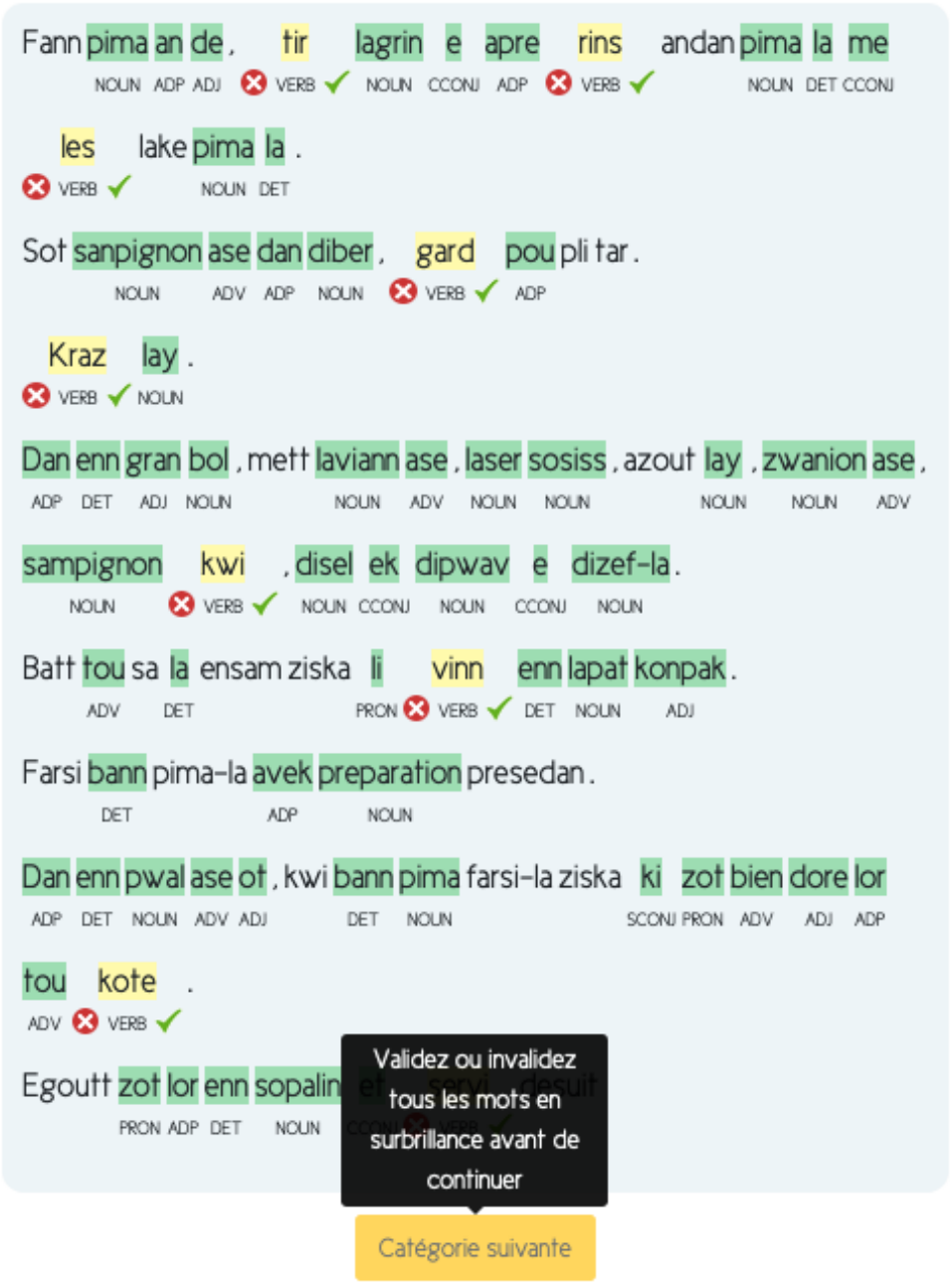


FIGURE 4.3 – Annotation de la catégorie VERB d'un texte.

Une fois les étiquettes validées ou invalidées, les participants ne peuvent pas vérifier leurs annotations, ni les modifier.

Ils peuvent néanmoins visualiser le guide d'annotation de toutes les catégories à tout moment en cliquant sur le menu déroulant à droite de l'écran d'annotation :

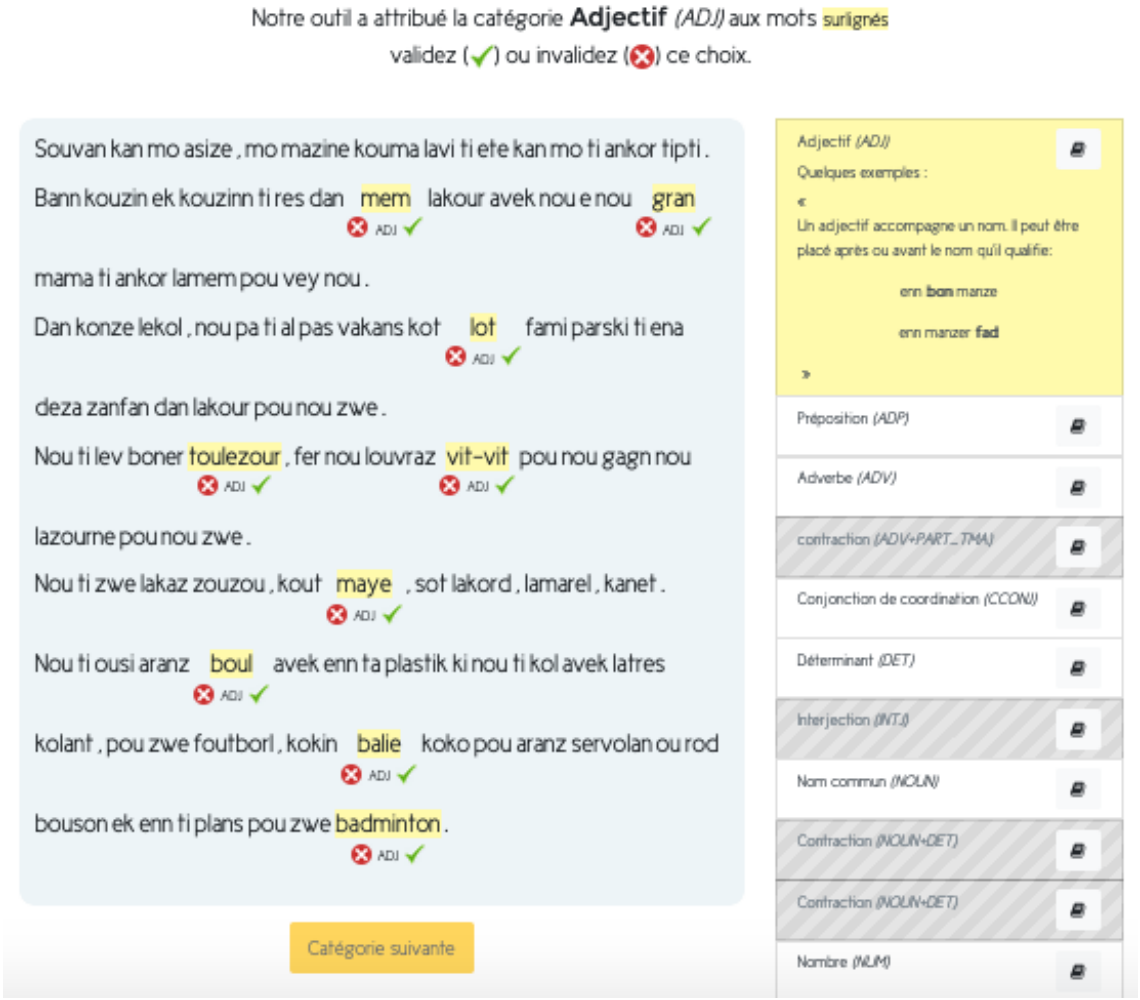


FIGURE 4.4 – Annotation de la catégorie ADJ et visualisation du guide d'annotation.

L'utilisateur n'est pas tenu de valider entièrement un texte, il peut s'arrêter à tout moment.

4.3 Résultats

4.3.1 La participation

Au 28 août 2019, 15 participants ont créé un compte, 9 ont déposé du texte et seulement 5 d'entre eux ont produit des annotations.

Nous rappelons que la plateforme a été lancé le 9 juillet 2019. Nous analysons les résultats obtenus pendant cette période de 50 jours – donc une durée relativement courte.

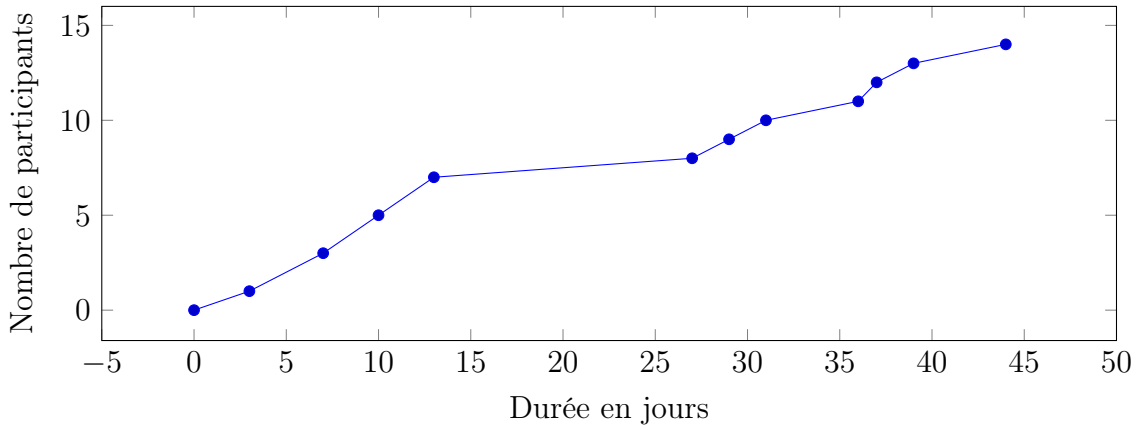


FIGURE 4.5 – L'évolution de la participation.

À partir du 25ème jour depuis le lancement, la participation est de nouveau en hausse après une période de stagnation. Cela coïncide avec la relance par mail des personnes rencontrées sur place.

En dépit de cela, peu de nouveaux participants se sont inscrit sur la plateforme et le taux de participation est resté décevant.

4.3.2 Le corpus et les annotations recueilli·es

Grâce à la contribution de ces participants, nous avons récolté 30 textes, soit 229 phrases pour 1 883 mots. Des 15 participant·e·s, 9 ont déposé du texte et 5 ont produit des annotations :

Type de texte	Nombre déposé	Nombre de tokens	Nombre d'annotations
Recettes	5	514	360
Poèmes	7	839	290
Proverbes	13	233	102
Texte libre	5	247	78
TOTAL	**30**	**1 883**	**830**

TABLE 4.4 – Nombre de textes et d'annotations produit·e·s par les participants

Sur les 830 tokens annotés par les utilisateurs, certains d'entre eux ont été attribués deux étiquettes différentes par des utilisateurs différents. Nous avons donc choisi une étiquette unique, nous basant sur le contexte dans lequel se trouvent ces tokens :

	Token	**MElt**	**Annotateur 1**	**Annotateur 2**	**Étiquette retenue**
1	mo	PRON	ADJ	PRON	PRON
2	mo	PRON	ADJ	PRON	PRON
3	mo	PRON	ADJ	PRON	PRON
4	ti	PART_TMA	PART_TMA	ADJ	ADJ
5	bann	DET	DET	NUM	DET
6	jeux	NOUN	X	NOUN	NOUN
7	bann	DET	DET	NUM	DET
8	losean	ADP	PROPN	NOUN	NOUN
9	indien	NOUN	PROPN	ADJ	NOUN
10	sa	PRON	PRON	DET	DET
11	divan	ADP	ADV	ADP	ADP
12	moris	PRON	PRON	PROPN	PROPN
13	deryer	ADV	ADV	ADP	ADP
14	pou	ADP	ADP	SCONJ	SCONJ

TABLE 4.5 – Tokens présentant différentes étiquettes.

Il est à noter qu'excluant les entrées 1, 2, 3, 6 et 12 de ce tableau, tous les tokens restants présentent une ambiguïté et peuvent se prêter à plus d'une étiquette.

4.3.3 La variation

Sur la plateforme, les utilisateurs peuvent proposer une orthographe alternative pour n'importe quel mot de la base de textes. Nous rappelons que bien qu'ayant connaissance qu'une graphie préconisée existe, une partie conséquente de la population ne la maîtrise pas. Ainsi, la variation que nous observons illustre l'appropriation du KM et de son écriture par ses locuteurs.

Assez paradoxalement, nous observons que pour certaines des entrées la variante proposée est son orthographe française [20].

Nous pouvons supposer qu'ils ne connaissent pas la graphie standard, et écrivent donc avec la graphie française pour ne pas commettre d'erreurs d'orthographe.

Nous avons également constaté que les mots écrits avec la graphie française ont bien été annoté avec une des étiquettes de partie du discours, et non avec l'étiquette X. Une seule occurrence étiquetée avec cette catégorie a été trouvée, ce qui démontre que les participants ont bien consulté le guide d'annotation.

20. Voir Annexe A.2.

4.3.4 Exactitude des modèles MElt entraînés

Du corpus entier récupéré sur *Ayo* (C_{Ayo}), nous avons extrait toutes les phrases annotées par MElt (C_{MElt}) et celles annotées uniquement par les utilisateurs ($C_{Utilisateurs}$).

Nous les avons extrait d'un fichier CSV contenant toutes les annotations recueillies sur Ayo avec une requête sur la base de données. Le fichier CSV obtenu a été manipulé avec Python pour récupérer C_{MElt} et $C_{Utilisateurs}$.

Les fichiers obtenus contenaient des textes en doublons, que nous avons retirés.

Ainsi, nous avons donc entraîné MElt sur le corpus de référence (C_{Ref}) et sur le corpus obtenu en couplant C_{Ref} et $C_{Utilisateurs}$:

Corpus d'entraînement	Nombre de phrases annotés	Nombre de tokens C_{Train}	Nombre de tokens C_{Test}	Exactitude
C_{Ref}	56	849	175	76 %
C_{Ref}+$C_{Utilisateurs}$	171	1 671	175	82 %

TABLE 4.6 – Entraînement de MElt.

Ces tests ont tous été effectués avec le lexique externe *Lalit*. Le même corpus de test (C_{Test}) est utilisé afin de pouvoir observer les éventuels impacts sur ce corpus quand nous modifions le corpus d'entraînement (C_{Train}).

Nous observons donc dans le tableau précédent (Table 4.6) une hausse de 6 points sur l'exactitude de l'outil de pré-annotation. Ayant un corpus peu conséquent, cela correspond seulement à 6 mots mieux annotés. De plus, en fusionnant $C_{Utilisateurs}$ et C_{Ref}, nous passons de 56 mots inconnus (mots qui n'apparaissent pas dans C_{Train}) par l'outil à 50 mots inconnus.

D'autres tests ont été effectué, mais étant peu concluants, nous avons choisi de ne pas les inclure dans ce rapport.

4.4 Discussion et évaluation des résultats

4.4.1 La qualité de l'annotation

L'annotation des utilisateurs

En manipulant $C_{Utilisateurs}$, nous avons observé que, compte tenu du contexte, un nombre non-négligeable de tokens semblaient être associés à une mauvaise étiquette. De ce fait la qualité de l'annotation produite par les utilisateurs de la plateforme serait relativement basse.

Nous pensons que l'absence d'une phase de formation à l'annotation a largement contribué à ce que les utilisateurs valident la pré-annotation sans la remettre en question.

Les annotations de MElt

L'exactitude des tests conduits sur le corpus de référence (C_{Ref}) est assez faible (76 %) [21] bien que nous pensons qu'elle aurait pu être meilleure si nous avions effectué l'apprentissage sur un corpus plus conséquent.

Dues aux contraintes de temps dans lequel ce travail est effectué, nous n'avons pas pu attribuer plus de temps à cette tâche. L'annotation manuelle du C_{Ref} ayant été faite par un seul annotateur, nous n'avions également pas les moyens d'annoter plus de textes manuellement dans les temps impartis.

Pour améliorer la qualité de la pré-annotation, nous pensions également entraîner MElt avec les variantes collectées mais nous n'avons malheureusement pas pu explorer cette piste.

4.4.2 Évaluation de la plateforme

Selon les retours des participants :

- la plateforme ne serait assez pas intuitive,
- le fait de devoir s'inscrire est rédhibitoire.

Nous pensons que l'adaptation de la plateforme pour collecter différents types de texte aurait compliqué l'interface d'accueil de la plateforme, et donc son utilisation. Nous n'avons pas passé suffisamment de temps à rendre l'interface plus attractive, faute de temps.

Il nous a également été suggéré de proposer une rémunération, même minime, pour la production d'annotation ou encore d'accepter que les participants nous envoient leurs textes afin que nous les mettions nous même sur la plateforme. Il est évident que ces procédés pour attirer plus de participants vont à l'encontre du principe même de la production participative. Nous n'avons donc pas intégré ces méthodes dans ce travail.

21. En comparaison, Krik (la plateforme pour le créole guadeloupéen) atteint 87 % d'exactitude lors de l'entraînement du corpus de référence (Karën Fort et Millour 2018).

4.4.3 Mise en perspective : Bisame, Krik et Ayo

	Bisame	Krik	Ayo
Outil de pré-annotation utlisé	Standford POS Tagger Melt	MElt	MElt
Type de textes traités	Recettes	Recettes	Recettes, poèmes, proverbes, textes libres
Sondage effectué Nombre de participants	Oui 1 224	Non	Oui 143
Nombre de personnes ayant créé un compte	208	35	15
Nombre de personnes ayant produit des annotations	47	11	5
Nombre de jours	109	9	50
Nombres d'annotations produites	24 588	1 205	1 024

TABLE 4.7 – Comparaison des trois plateformes. [22]

Comme nous pouvons l'observer dans le tableau ci-dessous, le nombre d'annotations produites sur Ayo est le plus bas des trois plateformes.

Cela peut être dû à plusieurs raisons :

- le manque de ressources pour une diffusion plus large de la campagne,
- une certaine confusion des locuteurs sur le but et les enjeux de plateforme,
- une crainte que le projet ne tienne pas sur la longeur, comme est le cas du projet Wikimédia pour le KM qui n'a pas été modifié depuis le mai 2017.

De plus, nous rappelons que la phase de formation n'a pas pu être intégré sur Ayo. Les participants se retrouvent donc face à la tâche d'annotation sans préparation préalable – qui a pu leur paraître insurmontable sans une connaissance pratique du jeu d'étiquettes.

Enfin, comparé à l'alsacien et au créole guadeloupéen, le KM est en voie d'être standardisé. Il se peut que le pourcentage élevé de la population mauricienne à ne pas avoir été initiée à l'écriture de la langue doutent de leur capacité à produire du contenu écrit en KM « correct ». Cela nous laisse à penser que lors de notre campagne de diffusion, nous n'avons pas assez mis l'accent sur le fait que nous traitons de toutes les représentations de la langue, qu'elles soient d'après le « standard » où non.

22. Les données de ce tableau pour l'alsacien et créole guadeloupéen ont été tirés de Karën FORT et MILLOUR 2018.

Conclusion

Ce projet nous a permis de comprendre la relation qu'entretenait les mauriciens avec leur langue, mais aussi de proposer un premier corpus annoté, librement disponible [23] pour le KM.

Le dialogue avec les locuteurs à travers le sondage que nous avons mené nous a permis de comprendre le besoin et les enjeux d'outiller une langue en voie de normalisation.

Malgré l'enthousiasme perçu à l'égard de notre travail, peu de locuteurs ont répondu à notre appel de participation. Bien que les résultats de se travail ne soient pas très concluants, nous osons espérer qu'avec un réaménagement de la plateforme ainsi qu'une meilleure stratégie pour sa diffusion, plus de gens reviendraient sur la plateforme.

Comme nous l'avons constaté, une différence de seulement 6 mots sur le corpus d'entraînement a augmenté l'exactitude de notre outil d'annotation de 76 % à 82 %. Cela démontre bien que tout l'intérêt de la méthodologie appliquée et également sa plus grande faiblesse : l'efficacité de notre outil dépend grandement du nombre de participants.

Une phase de formation obligatoire devra également être mise en place, afin que les annotations produites par les participants puissent être évaluées.

Enfin, nous avons appris qu'un correcteur en ligne (sujet à des droits d'auteurs) était en cours de production pour le KM et que depuis fin juillet 2019, une émission hebdomadaire est diffusé sur une des chaînes nationales pour initier la population à l'écriture et la lecture de la graphie standard.

Notre travail s'inscrit donc dans la même volonté de faire du KM une langue aussi résolument écrite que parlée, mais également de fournir des outils et des ressources numériques libres pour lui octroyer plus de visibilité sur Internet.

23. Le corpus annoté sera prochainement disponible sur la plateforme.

Annexes

Annexe A

A.1 Affiche pour la promotion de la plateforme sur les réseaux sociaux

FIGURE A.1 – Affiche pour la promotion de la plateforme.

A.2 Variantes récoltées

ID utilisateur	Variante 1	Variante 2	Variante 3
2	pouss	opuses	
1	zien	zuin	
1	ze	Ze	
1	lane	lannee	
2	ron	rond	
2	pou	pour	
2	leres diber	leress diber	
2	leres	leress	
2	tanperatir	temperatir	
2	pandan	pendan	
13	kouyer	kuyer	couyere
2	vie	vier	
2	karay	caraill	
1	kontign	contign	
1	les	less	
1	Angle	angle	
1	Franse	franse	
1	Kreol	kreol	
1	Per	per	
1	lapenn	la penn	
20	tranpe	trampe	
23	lake-zwanion	lake-zwanion	
1	Zien	zuin	
1	pe	p	
1	Ze	jeux	
1	ant	ent	
1	lane	lannee	
1	lane-la	lannee la	
7	bizin	bisin	
4	zwin	zoin	
4	nouri	nourri	
5	di riz	diri	
5	pwasson	poisson	
5	poi	pwa	
5	conzelé	konzele	
5	lapoud	la poud	

TABLE A.1 – Variantes proposées par les participants.

Bibliographie

Abney, Steven et Steven Bird (2010). « The human language project : building a Universal Corpus of the world's languages ». In : *Proceedings of the 48th annual meeting of the association for computational linguistics, juillet 2010*. (Upsal, Suède). Association for Computational Linguistics, p. 88–97.

Allen, Richard B (2008). « The constant demand of the French : the Mascarene slave trade and the worlds of the Indian Ocean and Atlantic during the eighteenth and nineteenth centuries ». In : *The Journal of African History* 49.1, p. 43–72.

Atchia-Emmerich, Bilkiss (2005). « La situation linguistique à l'ıle Maurice. Les développements récents à la lumière d une enquête empirique ». In : *Université de Nuremberg, dissertation inaugurale à la Faculté de philosophie II (science de la langue et de la littérature, 20 janvier 2005*.

Baggioni, Daniel et Didier de Robillard (1990). *Ile Maurice : une francophonie paradoxale*. Editions L'Harmattan.

Baker, Philip (1972). *Kreol : a description of Mauritian Creole*. C. Hurst, Londres.

— (2007). « Elements for a sociolinguistic history of Mauritius and its creole (to 1968) ». In : *The making of Mauritian creole*, p. 307–333.

Baker, Philip et Chris Corne (1982). *Isle de France creole : affinities and origins*. Karoma Ann Arbor.

Baker, Philip et Vinesh Y Hookoomsing (1987). *Diksyoner kreol morisyen : Dictionary of Mauritian Creole*. L'Harmattan.

Baker, Philip et Sibylle Kriegel (2013). « Mauritian Creole ». In : *The survey of pidgin and creole languages. Volume 2 : Portuguese-based, Spanish-based, and French-based Languages*. Sous la dir. de Susanne Maria Michaelis, Philippe Maurer, Martin Haspelmath et Magnus Huber. Oxford : Oxford University Press. url : `https://apics-online.info/surveys/55`.

Baker, Philip et Anand Syea (1991). « On the copula in Mauritian Creole, past and present ». In : *T. Byrne, F. & Huebner (ed.), Development and Structures in Creole Languages*, p. 159–175.

Barbaresi, Adrien (2013). « Challenges in web corpus construction for low-resource languages in a post-BootCaT world ». In : *6th Language & Technology Conference, Less Resourced Languages special track, décembre 2013*. (Poznan, Polande), p. 69–73. url : `https://halshs.archives-ouvertes.fr/halshs-00919410`.

Barteld, Fabian (2017). « Detecting spelling variants in non-standard texts ». In : *Proceedings of the student research workshop at the 15th conference of the*

European chapter of the association for computational linguistics, EACL 2017, 3-7 avril. (Valence, Espagne), p. 11–22.

Biemann, Chris, Gerhard Heyer, Uwe Quasthoff et Matthias Richter (2007). « The Leipzig Corpora Collection-monolingual corpora of standard size ». In : *Proceedings of Corpus Linguistic* 2007.

Böhmová, Alena, Jan Hajič, Eva Hajičová et Barbora Hladká (2003). « The Prague dependency treebank ». In : *Treebanks*. Springer, p. 103–127.

Bonami, Olivier et Fabiola Henri (2010). « How complex is Creole Inflectional Morphology ? » In : *Budapest : International Meeting of Morphology, 7 juillet*. (Budapest, Hongrie).

Burger-Helmchen, Thierry et Julien Pénin (2011). « Crowdsourcing : définition, enjeux, typologie ». In : *Management Avenir* 1, p. 254–269.

Carpooran, Arnaud (2011). *Lortograf Kreol Morisien*. Akademi Kreol Morisien.

Chaudenson, Robert (1979). « Créoles français de l'Océan Indien et langues africaines ». In : *Hancock, Ian F. éd. : Readings in Creole Studies, Gent : Story-Scientia*, p. 217–237.

— (2013). « Approche (historico-) linguistique des créoles des Mascareignes et des Seychelles ». In : *Études océan Indien* 49-50.

Dabre, Raj, Aneerav Sukhoo et Pushpak Bhattacharyya (2014). « Anou Tradir : Experiences In Building Statistical Machine Translation Systems For Mauritian Languages - Creole, English, French ». In : *Proceedings of the 11th International Conference on Natural Language Processing, ICON 2014, December 18-21, 2014*. (Goa, Inde), p. 82–88.

DeGraff, Michel (2005a). « Do Creole languages constitute an exceptional typological class ? » In : *Revue française de linguistique appliquée* 10.1, p. 11–24.

— (2005b). « Linguists' most dangerous myth : The fallacy of Creole Exceptionalism ». In : *Language in society* 34.4, p. 533–591.

Denis, Pascal et Benoıt Sagot (2009). « Coupling an annotated corpus and a morphosyntactic lexicon for state-of-the-art POS tagging with less human effort ». In : *Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation, Volume 1, 3-5 décembre*. (Hong Kong, Chine).

Durrleman, Stephanie (2014). « The Syntax of Mauritian Creole Anand Syea, Bloomsbury Studies in Theoretical Linguistics (2013) ». In : *Lingua* 145, p. 226–242.

Eriksen, Thomas Hylland (1990). « Linguistic diversity and the quest for national identity : The case of Mauritius ». In : *Ethnic and racial studies* 13.1, p. 1–24.

— (2018). « Language and Ethnic Hierarchy in Mauritius ». In : *Creolization and Pidginization in Contexts of Postcolonial Diversity*. Leiden, Pays-Bas : Brill, p. 59. isbn : 9789004363397. url : `https://brill.com/view/book/edcoll/9789004363397/BP000013.xml`.

Fattier, Dominique (2003). « Grammaticalisations en créole haıtien : morceaux choisis ». In : *Creolica* 13, p. 59.

Ferguson, Charles (1959). « Diglossia ». In : *word* 15.2, p. 325–340.

FORT, Karën et Alice MILLOUR (2018). « À l'écoute des locuteurs : production participative de ressources langagières pour des langues non standardisées ». In : *Traitement Automatique des Langues*, p. 59–3.

FORT, Karën et Benoıt SAGOT (2010). « Influence of pre-annotation on POS-tagged corpus development ». In : *Proceedings of the Fourth Linguistic Annotation Workshop, juillet 2010*. Upsal, Suède : Association for Computational Linguistics, p. 56–63. URL : https://www.aclweb.org/anthology/W10-1807.

GARRETTE, Dan, Jason MIELENS et Jason BALDRIDGE (2013). « Real-World Semi-Supervised Learning of POS-Taggers for Low-Resource Languages ». In : *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL 2013, 4-9 août, Volume 1 : Long Papers*. (Sofia, Bulgarie), p. 583–592.

GONZALEZ, Paola Carrión et Emmanuel CARTIER (2012). « Technological tools for dictionary and corpora building for minority languages : example of the French-based Creoles ». In : *Language Technology for Normalisation of Less-Resourced Languages*, p. 47.

GOODCHILD, Samantha (2013). « Being Mauritian : A Sociolinguistic Case Study on the Transmission and Use of Mauritian Creole in the UK ». In : *Newcastle Working Papers in Linguistics* 19.1, p. 109–127.

GOVERNMENT OF MAURITIUS (2011). *Housing and Population Census, Volume II : Demographic and fertility characteristics*. URL : http://statsmauritius.govmu.org/English/Documents/publications/Housing/economics%5C%20and%5C%20social%5C%20indicators/reports/2011VolIIPC.pdf.

— (p.d.). *Explore Mauritius - History*. URL : http://www.govmu.org/English/ExploreMauritius/Pages/History.aspx.

GRANT, Anthony (2009). « Admixture, Structural Transmission, Simplicity and Complexity ». In : *Simplicity and Complexity in Creoles and Pidgins*. Battlebridge Publications, p. 125–152.

HARMON, Jimmy (2011). « Le système éducatif de l'Île Maurice ». In : *Revue internationale d'éducation de Sèvres,* 57, p. 22–30.

HAZAËL-MASSIEUX, Marie-Christine (1999). *Les créoles : l'indispensable survie*. Paris : Éditions Entente, coll. « Langues en péril », p. 15.

— (2002). « Les créoles à base française : une introduction ». In : *Travaux Interdisciplinaires du Laboratoire Parole et Langage d'Aix-en-Provence (TIPA)* N° 21, p. 63–86.

— (2005). « L'écriture des créoles français au début du 3e millénaire : état de la question ». In : *Revue française de linguistique appliquée* 10.1, p. 77–90.

HENRI, Fabiola et Olivier BONAMI (2016). « Prédire l'agglutination de l'article en mauricien ». In : *Faits de langue*.

HENRI, Fabiola, Gregory STUMP et Delphine TRIBOUT (2017). *Conversion relations and the morphologicalcomplexity of three French-based creoles*.

HOOKOOMSING, Vinesh Y (2004). *A Harmonized Writing System for the Mauritian Creole Language : Grafi-Larmoni*.

KING, Benjamin Philip (2015). *Practical Natural Language Processing for Low-Resource Languages.*

KRAUSS, Michael (1992). « The world's languages in crisis ». In : *Language* 68.1, p. 4–10.

LECLERC, Jacques (2001). *Créoles, L'aménagement linguistique dans le monde.* Québec, TLFQ–Université Laval. URL : http://www.axl.cefan.ulaval.ca/amsudant/creole.htm.

— (p.d.). *L'aménagement linguistique dans le monde : Île Maurice.* Québec, TLFQ–Université Laval. URL : http://www.axl.cefan.ulaval.ca/afrique/maurice.htm. Mise à jour : 23 janvier 2007.

LEWIS, Will (2010). « Haitian Creole : How to Build and Ship an MT Engine from Scratch in 4 days, 17 hours, & 30 minutes ». In : *EAMT 2010 : Proceedings of the 14th Annual conference of the European Association for Machine Translation. 8pp, 27–28 mai.* (Saint-Raphaël, France).

MAMODE, Mei-Lan (2013). « À l'encontre des présupposés linguistiques : morphologie flexionnelle et dérivationnelle du créole mauricien ». In : *Bilingual Workshop on Theoretical Linguistics, décembre 2013.* (Waterloo, Ontario, Canada).

MASON, Marilyn (2000). « Issues from Corpus Analysis that have influenced the On-going Development of Various Haitian Creole Text- and Speech-based NLP Systems and Applications ». In : *Proceedings of the Second International Conference on Language Resources and Evaluation, LREC 2000, 31 mai - 2 juin, 2000.* (Athènes, Grèce). URL : http://www.lrec-conf.org/proceedings/lrec2000/pdf/342.pdf.

MCDONALD, Ryan, Joakim NIVRE, Yvonne QUIRMBACH-BRUNDAGE, Yoav GOLDBERG, Dipanjan DAS, Kuzman GANCHEV, Keith HALL, Slav PETROV, Hao ZHANG, T OSCAR et al. (2013). « Universal dependency annotation for multilingual parsing ». In : *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2 : Short Papers), août 2013.* (Sofia, Bulgarie), p. 92–97. URL : https://www.aclweb.org/anthology/P13-2017.

MCWHORTER, John (2001). « The world's simplest grammars are creole grammars ». In : *Linguistic typology* 5.2, p. 125–66.

MICHAELIS, Suzanne Maria, Phillipe MAURER, Martin HASPELMATH et Magnus HUBER (2013). *APiCS online. Max Planck Institute for Evolutionary Anthropology, Leipzig.*

MILLOUR, Alice (2019). « Getting to Know the Speakers : a Survey of a Non-Standardized Language Digital Use ». In : *9th Language & Technology Conference : Human Language Technologies as a Challenge for Computer Science and Linguistics, mai 2019.* (Poznań, Poland). URL : https://hal.archives-ouvertes.fr/hal-02137280.

MILLOUR, Alice et Karën FORT (2016). *Construction de ressources langagières annotées par myriadisation pour le traitement automatique des langues peu dotées : le cas de l'alsacien.* Rapp. tech. 113.

— (2018). « Krik : First Steps into Crowdsourcing POS tags for Kréyòl Gwadloupéyen ». In : *Collaboration and Computing for Under-Resourced Languages*

(CCURL), mai 2018. (Miyazaki, Japon). URL : https://hal.archives-ouvertes.fr/hal-01790617.

Millour, Alice et Karën Fort (2018). « Toward a Lightweight Solution for Less-resourced Languages : Creating a POS Tagger for Alsatian Using Voluntary Crowdsourcing ». In : *Eleventh International Conference on Language Resources and Evaluation (LREC), mai 2018*. (Miyazaki, Japon). URL : https://hal.archives-ouvertes.fr/hal-01790615.

Muysken, Pieter, Norval Smith et al. (1995). « The study of pidgin and creole languages ». In : *Pidgins and creoles : An introduction*, p. 3–14.

Nye, David Edwin (2008). *Technologie & Civilisation : 10 questions fondamentales liées aux technologies*. Fyp éditions, p. 88.

Petrov, Slav, Dipanjan Das et Ryan McDonald (2011). « A universal part-of-speech tagset ». In : *arXiv preprint arXiv :1104.2086*.

Portail touristique de l'Île Maurice (p.d.). *Histoire de l'Ile Maurice*. URL : https://www.ile-maurice.fr/infos-pratiques/histoire-et-geographie/histoire-de-l-ile-maurice.html.

Rajah-Carrim, Aaliya (2009). « Use and standardisation of Mauritian Creole in electronically mediated communication ». In : *Journal of Computer-Mediated Communication* 14.3, p. 484–508.

Sagot, Benoît (2016). « Multilingual part-of-speech tagging with MElt ». In : *23ème Conférence sur le Traitement Automatique des Langues Naturelles, juillet 2010*. (Paris, France). URL : https://hal.inria.fr/hal-01352243.

Sagot, Benoıt, Karën Fort, Adda Gilles, Joseph Mariani et Bernard Lang (2011). « Un turc mécanique pour les ressources linguistiques : critique de la myriadisation du travail parcellisé ». In : *TALN'2011 - Traitement Automatique des Langues Naturelles, juin 2011*. (Montpellier, France). URL : https://hal.inria.fr/inria-00617067.

Scannell, Kevin P (2007). « The Crúbadán Project : Corpus building for under-resourced languages ». In : *Building and Exploring Web Corpora : Proceedings of the 3rd Web as Corpus Workshop, janvier 2007*. (Louvain-la-Neuve, Belgique). T. 4, p. 5–15.

Thomason, Sarah Grey et Terrence Kaufman (1991). *Language Contact, Creolization, and Genetic Linguistics*. Univ of California Press.

United Nations (p.d.). *World Population Prospects : The 2017 Revision*. URL : https://population.un.org/wpp/DataQuery/.