Outiller la documentation des langues créoles

Eric Le Ferrand, Claudel Pierre-Louis, Ruoran Dong, Benjamin Lecouteux, Daphné Gonçalves-Teixeira, William N Havard, Emmanuel Schang

Outiller la documentation des langues créoles

Éric Le Ferrand[1,2], Claudel Pierre-Louis[1], Ruoran Dong[3], Benjamin Lecouteux[3], Daphne Gonçalves-Teixeira[1], William N. Havard[1], Emmanuel Schang[1]
(1) LLL, 10 rue de Tours, BP 46527 - 45072 ORLEANS CEDEX 2 (FRANCE)
(2) Boston College, Boston (USA)
(3) LIG, Université Grenoble Alpes (FRANCE)
william.havard@univ-orleans.fr, leferran@bc.edu,
daphne.goncalves-teixeira@univ-orleans.fr,
benjamin.lecouteux@univ-grenoble-alpes.fr,
emmanuel.schang@univ-orleans.fr

RÉSUMÉ

Ce papier propose donc un retour d'expérience basé sur l'emploi d'outils informatiques sur différents terrains linguistiques concernant les langues créoles.

ABSTRACT

Tooling up Creole Languages Documentation

This paper provides feedback based on the use of IT tools in different linguistic fields concerning Creole languages.

MOTS-CLÉS : Documentation des langues, créoles, traitement automatique de la parole.

KEYWORDS: Language Documentation, Creole Languages, Automatic Speech Processing.

1 Introduction

Dans ce papier, nous passons en revue les pistes suivies par le projet CREAM (documentation des langues CREoles Assistée par la Machine, ANR CS38) pour la documentation outillée des langues créoles. Afin de répondre aux besoins distincts de ses utilisateurs, principalement des linguistes, plusieurs solutions sont présentées.

La documentation des langues consiste à mettre à disposition des données langagières sous une forme consultable par ses utilisateurs [1].

> Documentation of a language is an activity (and, derivatively, its result) that gathers, processes and exhibits a sample of data of the language that is representative of its linguistic structure and gives a fair impression of how and for what purposes the language is used. (Lehmann, 2001)

1. Voir également : (Austin, 2016; Michaud *et al.*, 2016) entre autres.

Il est essentiel de noter que les langues créoles ne partagent pas les mêmes niveaux d'écriture et de standardisation. Certaines langues jouissent d'un statut officiel, à l'instar du santoméen à São Tomé et Príncipe, l'une des langues nationales de l'archipel. D'autres, comme le créole de l'île Maurice, sont parlées par la majorité de la population sans bénéficier d'un statut officiel reconnu.

Alors qu'il est généralement admis que le Traitement Automatique des Langues (TAL) vise principalement à faciliter la graphisation (Chaudenson, 2005) des langues minoritaires, notre approche diverge en ne conférant pas à la graphisation un rôle prédominant au sein de ce projet. En effet, les résultats de nos travaux sur le terrain nous orientent vers l'utilisation des outils du TAL à des fins autres que la transcription.

Ainsi, cet article apporte un retour d'expérience sur l'application d'outils informatiques dans divers contextes linguistiques en milieu créolophone.

2 Transcription automatisée

Le besoin de transcription automatique et les modalités de sa réalisation sont intrinsèquement liés à l'existence d'une norme orthographique stable et acceptée. Pour les langues créoles qui disposent d'une orthographe normalisée, comme le créole de la Guadeloupe (kréyòl gwadloupéyen), la transcription automatique est d'une aide précieuse pour le linguiste. Nous avons reçu plusieurs demandes de transcription automatique émanant de linguistes locaux ou de projets de recherche récoltant des données de terrain.

2.1 Transcrire automatiquement depuis le terrain

Au cours d'un travail de terrain en Guadeloupe effectué par l'un des auteurs en soutien à un projet de recherche (NSF-IRES 1952568 : Experimental linguistics in the Caribbean), un ensemble d'enregistrements de terrain nous a été fourni avec pour seule directive de fournir des transcriptions générées automatiquement. Ces enregistrements présentaient une grande diversité en termes de nature et de qualité. Certains comprenaient une série de jugements grammaticaux en français, tandis que d'autres contenaient des discussions en anglais sur des phénomènes linguistiques. La plupart d'entre eux, cependant, renfermaient de la parole spontanée en créole dans divers contextes tels que des visites guidées, des séminaires et des conversations.

Notre sélection a porté sur un enregistrement d'une heure renfermant une quantité suffisante de contenu en créole mêlé à des segments entièrement en français, des segments où la syntaxe créole et française se mélangent (*code-switching*) et une poignée de segments en anglais. Dans un premier temps, nous avons appliqué un algorithme de détection d'activité vocale

(VAD)[2] pour identifier les segments de parole. À cette fin, nous disposions d'un modèle de reconnaissance de la parole préalablement entraîné sur une heure de parole transcrite (Macaire *et al.*, 2022).

Ce modèle repose sur l'architecture WAV2VEC2 (Baevski *et al.*, 2020), une architecture de transformers utilisant un apprentissage auto-supervisé pour extraire les paramètres acoustiques. Dans notre cas spécifique, le modèle LeBenchmark a été employé (Evain *et al.*, 2021). La représentation générée par cette architecture est ensuite alimentée à une tête de transformers entraînée sur un corpus de parole transcrite, utilisant une fonction CTC (*Connectionist Temporal Classification*, Graves *et al.* 2006). Cette fonction CTC apprend à générer la probabilité d'un caractère pour chaque paramètre acoustique.

Le processus se poursuit avec un décodage final modulé à l'aide d'un modèle de langue, lui-même entraîné sur les mêmes transcriptions qui ont servi à l'entraînement initial du modèle. Cette approche assure une cohérence et une précision accrues dans la transcription automatique des enregistrements en créole, en tirant parti de la représentation riche des caractéristiques acoustiques acquises par l'architecture WAV2VEC2, tout en affinant le résultat final à l'aide d'un modèle de langue formé sur le corpus spécifique utilisé pour l'entraînement initial.

Ainsi, nous avons appliqué le modèle développé sur les segments détectés par la VAD. Un Gold standard, établi en se basant sur nos propres transcriptions, nous a été fourni, permettant ainsi une évaluation concrète des performances du modèle. Les résultats obtenus révèlent une performance relativement faible avec un taux d'erreur par mot (WER) à 73% et un taux d'erreur par caractère (CER) à 45%. Cependant, il est important de noter que malgré ces limitations, l'utilisation de ce modèle offre aux linguistes de terrain un gain de temps considérable, variable toutefois en fonction de la qualité de l'audio.

Parmi les retours que nous avons recueillis de la part des linguistes de terrain, la détection des segments de parole a été particulièrement saluée pour le temps qu'elle permet d'économiser. De plus, bien que notre évaluation semble indiquer une performance modérée du système de transcription, il faut souligner que la qualité des transcriptions varie considérablement au sein d'un même enregistrement. Une transcription générée peut ainsi présenter une faible qualité pour des segments très bruités, tandis que d'autres nécessiteront uniquement des corrections mineures. Cette hétérogénéité souligne la complexité de la tâche de transcription automatique, tout en mettant en lumière les avantages substantiels que notre approche peut offrir dans le contexte spécifique des langues créoles.

2. https://github.com/amsehili/auditok

Analyse d'erreur	Référence	Transcription automatique
La nasale finale est reconnue comme deux orales	zot matinike gwadloupeyen	zoln patinike gwadloup ee
la référence est en français	deux saisons	deu sezon
erreur de segmentation	zo kay an grante	jo kay angrandte
erreurs de segmentation et de transcription	byen pale de bonda nou kay soukre bonda	mye fame de gonda nou ka ai soucebo

TABLE 1 – Examples de transcriptions

2.2 Aligner les transcriptions sur la parole

Par ailleurs, en appliquant les mêmes techniques, nous avons procédé à l'alignement d'un corpus audio en créole haïtien. Ce corpus, composé de 10 enregistrements audio associés à des transcriptions au format MSWord, présentait la particularité de ne pas comporter de balises temporelles facilitant l'alignement. Pour résoudre cette problématique, nous avons opté pour une transcription automatique de type 'gros grain', générée à partir de 10 minutes de parole préalablement alignée, et traitée ultérieurement à l'aide de WAV2VEC2.

Cette approche a permis d'obtenir une représentation plus robuste du contenu audio, malgré l'absence initiale de balises temporelles. L'utilisation de la transcription automatique 'gros grain' a servi de prétraitement, suivie d'une étape cruciale impliquant le modèle WAV2VEC2 pour affiner et préciser l'alignement temporel. Cette démarche souligne l'adaptabilité de notre méthode dans des contextes divers, renforçant ainsi sa pertinence pour des langues créoles variées et des corpus audio présentant des défis spécifiques.

3 TètKole et l'interprétation

TètKole, signifiant 'ensemble' en créole haïtien, représente un outil conçu dans le cadre du projet CREAM par des étudiants en informatique de l'université de Chambéry. Il s'agit d'un redéveloppement de l'outil LIG-AIKUMA, qui était précédemment disponible sur la plateforme Android (Blachon *et al.*, 2016). TètKole a pour fonction de permettre l'interprétation (traduction à l'oral) d'enregistrements réalisés par des linguistes de terrain. Il vise à l'obtention de corpus de parole parallèles dans un objectif de documentation des langues.

L'utilisation de LIG-AIKUMA dans notre projet s'est heurtée à trois obstacles majeurs, justifiant ainsi le développement de TètKole :

Stabilité sur tablette et téléphone : L'instabilité de son fonctionnement sur tablette ou téléphone a été un point de friction significatif, suscitant des retours négatifs de l'ensemble des utilisateurs.

Préférence des linguistes pour les PC : Les linguistes du projet manifestent une préférence marquée pour travailler sur PC plutôt que sur téléphone ou tablette, constituants initiaux de LIG-AIKUMA. Cette préférence a motivé le besoin d'une version mieux adaptée aux environnements de travail privilégiés par les linguistes.

Limite de durée des fichiers son : Une autre limitation était imposée par la contrainte des téléphones portables et des tablettes, qui ne permettaient pas de travailler sur des fichiers son d'une durée dépassant les 15 minutes. Cette restriction ne répondait pas aux exigences des linguistes du projet, qui avaient exprimé le besoin de manipuler des enregistrements de durées plus étendues.

En réponse à ces limitations, TètKole a été développé pour surmonter ces obstacles spécifiques et répondre de manière plus adéquate aux besoins et préférences des linguistes du projet CREAM.

TètKole est donc un redéveloppement en Java de certaines fonctionnalités de LIG-AIKUMA. L'outil est librement disponible sur https://github.com/LLL-Orleans/TetKole sous licence GPL. Il a été réalisé par des étudiants de Master Informatique de l'université de Savoie Mont Blanc (Chambéry) dans le cadre d'un projet.

TètKole offre une interface conviviale pour les linguistes, leur permettant de charger un enregistrement au format .wav ou .mp3. À l'aide de la souris, le linguiste peut sélectionner avec précision la portion de parole à interpréter, comme illustré par les barres verticales sur la Figure 1. En suivant cette sélection, en cliquant sur l'icône dédiée à l'enregistrement, le linguiste génère un fichier son contenant son interprétation.

Cette approche centrée sur l'interaction visuelle et l'ergonomie vise à simplifier le processus d'interprétation, offrant aux linguistes un moyen intuitif et efficace pour traiter les enregistrements et produire des interprétations orales. La combinaison de la sélection visuelle et de la création rapide de fichiers son facilite le flux de travail des linguistes, favorisant ainsi une utilisation efficace de l'outil TètKole dans le cadre du projet CREAM.

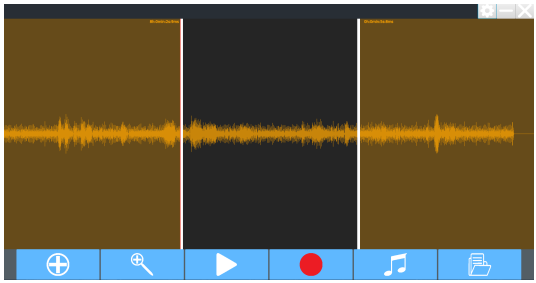

FIGURE 1 – L'écran d'accueil de TètKole après chargement du fichier son à interpréter

Le linguiste peut vérifier la qualité de son travail et reprendre ses interprétations (supprimer les interprétations non satisfaisantes) ou les valider (Figure 2).

Cet outil a permis la réalisation d'un corpus oral parallèle français /haïtien d'environ 3 heures. A partir de la version française, il est aisé d'obtenir une transcription du corpus en français avec Whisper un modèle de transcription automatique développé par OpenAI. On dispose

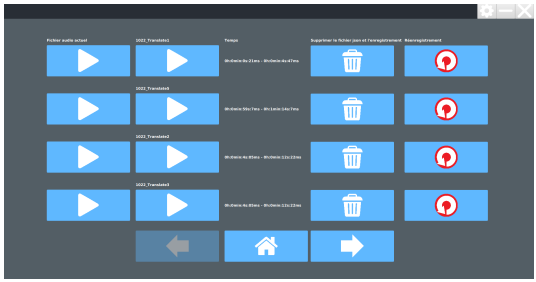

FIGURE 2 – L'écran de validation des interprétation de TètKole

alors de la parole en haïtien, de la parole en français et de la transcription en français.

4 Perspectives Futures

4.1 Valorisation des enregistrements de terrain

Des années de travail de terrain par une multitude de linguistes ont permis de collecter un grand nombre d'enregistrements, dont seule une petite fraction a été transcrite. Ces enregistrements n'ont généralement pas été valorisés par le passé, faute d'une solution technique appropriée. L'avénement des modèles d'apprentissage auto-supervisé (*self-supervised learning*, SSL) qui ne nécessitent pas de transcriptions permettent désormais d'envisager une valorisation de ces données.

Nous souhaitons donc étudier l'utilisabilité des données de terrain *déjà existantes* dans cadre d'un pré-entraînement auto-supervisé de modèles WAV2VEC2. L'utilisation de ces données représente un défi, et ce à deux titres. D'une part, ces données ont généralement été enregistrées sur cassettes, puis digitalisées, et sont donc de qualité acoustique variable. D'autre part, ces enregistrements font suite à des enquêtes de terrain répondant à des problèmatiques scientifiques diverses (par exemple des enquêtes grammaticales, atlas sonore, etc.) et ne rentrent pas dans le canon standard des données actuellement utilisées pour entraîner de tel modèles (généralement des livres audios) contenant en grande majorité de parole lue.

Ainsi nous souhaitons explorer l'influence de plusieurs variables sur la qualité des modèles entraînés. Plus précisément, nous explorerons l'influence de la nature des données de terrain, de la qualité acoustique, du ratio parole lue/parole spontanée, et de la quantité de données minimale nécessaire. Ces questions restent à ce jour encore ouvertes, surtout dans le cadre des langues peu dotées.

4.2 Des modèles neuronaux et des questionnements linguistiques

Les modèles neuronaux que nous entraînerons seront mis au service d'un questionnement linguistique. Ainsi, ces modèles nous permettront d'explorer l'influence des langues lexificatrices [3] sur les créoles. Pour explorer cette question, nous souhaitons donc entraîner des modèles auto-supervisés "pan-créoles" (incluant par exemple des données d'haïtien, de martiniquais, de réunionais, etc.) à partir de rien (*from scratch*), à partir de leur langue lexificatrice, à partir d'une autre langue, ou bien à partir de plusieurs langues (en utilisant des modèles multilingues, de type XLSR (Conneau *et al.*, 2021)). Cela nous permettra ainsi de voir si les créoles partagent plus de traits entre-eux qu'avec leur langue lexificatrice, auquel cas le modèle pan-créole se révélerait meilleur une fois raffiné (*fine-tuned*) sur chaque créole pris indépendamment, ou non. Les résultats de Macaire *et al.* (2022) semblent indiquer qu'il est souhaitable d'entraîner des modèles de reconnaissance de parole à partir de la langue lexificatrice plutôt que de modèles multilingues. Cependant, cela reste à vérifier sur des créoles ayant une autre langue lexificatrice.

Références

AUSTIN P. K. (2016). Language documentation 20 years on. *Endangered languages and languages in danger : Issues of documentation, policy, and language rights*, p. 147–170.

BAEVSKI A., ZHOU Y., MOHAMED A. & AULI M. (2020). wav2vec 2.0 : A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, **33**, 12449–12460.

BLACHON D., GAUTHIER E., BESACIER L., KOUARATA G.-N., ADDA-DECKER M. & RIALLAND A. (2016). Parallel speech collection for under-resourced language studies using the lig-aikuma mobile device app. *Procedia Computer Science*, **81**, 61–66.

CHAUDENSON R. (2005). Description et graphisation : le cas des créoles français. *Revue française de linguistique appliquée*, **10**(1), 91–102.

CONNEAU A., BAEVSKI A., COLLOBERT R., MOHAMED A. & AULI M. (2021). Unsupervised Cross-Lingual Representation Learning for Speech Recognition. In *Proc. Interspeech 2021*, p. 2426–2430. DOI : 10.21437/Interspeech.2021-329.

EVAIN S., NGUYEN H., LE H., BOITO M. Z., MDHAFFAR S., ALISAMIR S., TONG Z., TOMASHENKO N., DINARELLI M., PARCOLLET T. *et al.* (2021). Lebenchmark : A reproducible framework for assessing self-supervised representation learning from speech. In *INTERSPEECH 2021 : Conference of the International Speech Communication Association*.

3. On désigne par *langue lexificatrice*, les langues "dont les créoles ont retenu la plus grande partie de leur lexique si ce n'est de leur grammaire" (Mufwene, 2021)

GRAVES A., FERNÁNDEZ S., GOMEZ F. & SCHMIDHUBER J. (2006). Connectionist temporal classification : Labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, p. 369–376, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/1143844.1143891.

LEHMANN C. (2001). *Language documentation : A program*. na.

MACAIRE C., SCHWAB D., LECOUTEUX B. & SCHANG E. (2022). Automatic speech recognition and query by example for creole languages documentation. In *Findings of the Association for Computational Linguistics : ACL 2022*, p. 2512–2520.

MICHAUD A., GUILLAUME S., JACQUES G., MAC D.-K., JACOBSON M., PHAM T.-H. & DEO M. (2016). Contribuer au progrès solidaire des recherches et de la documentation : la collection pangloss et la collection auco. In *Journées d'Etude de la Parole 2016*, volume 1, p. 155–163.

MUFWENE S. (2021). Créoles. *Langage et société*, **Hors série**(HS1), 81–86. DOI : 10.3917/ls.hs01.0082.