Modèles auto-supervisés de traitement de la parole pour le Créole Haitien

William N Havard, Renauld Govain, Benjamin Lecouteux, Emmanuel Schang

Modèles auto-supervisés de traitement de la parole pour le Créole Haitien

William N. Havard[1, 2] Renauld Govain[3]
Benjamin Lecouteux[2] Emmanuel Schang[1]
(1) LLL, Université d'Orléans, CNRS, F-45000 Orléans, France
(2) LIG, Université Grenoble Alpes, CNRS, Grenoble INP, F-38000 Grenoble, France
(3) LangSé, Université d'État d'Haïti, Port-au-Prince, Haïti
william.havard@univ-orleans.fr

RÉSUMÉ

Nous développons des modèles de traitement de la parole spécifiquement dédiés au créole haïtien (*kreyòl*), le positionnant ainsi comme une langue bien dotée en termes de modèles auto-supervisés de traitement de la parole. Pour ce faire, nous pré-entraînons des modèles monolingues WAV2VEC2-BASE, WAV2VEC2-LARGE et DATA2VEC-AUDIO-BASE à partir de zéro, qui sont ensuite affinés pour une tâche de reconnaissance automatique de la parole. Nous comparons la performance de ces modèles avec des modèles affinés à partir de modèles multilingues (XLSR-53, XLSR2-300M, MMS-1B) et monolingues basés sur le français (LEBENCHMARK 1 à 7K). Nos résultats démontrent l'efficacité du pré-entraînement monolingue, avec des performances pouvant rivaliser, voire surpasser, celle de grands modèles multilingues. Ce travail propose ainsi des modèles robustes de reconnaissance vocale pour le *kreyòl*, adaptables à d'autres créoles français des Caraïbes, contribuant ainsi au développement technologique de ces langues peu dotées.

ABSTRACT

Self-Supervised Models of Speech Processing and Recognition for Haitian Creole

We develop customised speech processing models for Haitian Creole (*Kreyòl*), positioning it as a high-resource language in terms of self-supervised models of speech processing. To do so, we pre-train monolingual WAV2VEC2-BASE, WAV2VEC2-LARGE and DATA2VEC-AUDIO-BASE models from scratch, which are then finetuned on an automatic speech recognition task. We compare the performance of these models with models finetuned from multilingual (XLSR-53, XLSR2-300M, MMS-1B) and monolingual French-based (LEBENCHMARK 1 à 7K) models. Our results demonstrate the effectiveness of monolingual pre-training, with performance that can rival or even surpass that of large multilingual models. This work thus proposes robust speech transcription models for *kreyòl*, adaptable to other Caribbean French creoles, contributing to the technological development of these under-resourced languages.

MOTS-CLÉS : créole haïtien, modèles auto-supervisés, traitement de la parole.

KEYWORDS: Haitian Creole, self-supervised models, speech processing.

ARTICLE : **Soumis à** Interspeech 2025.

1 Introduction

Le créole haïtien (*kreyòl*, ou *kreyòl ayisyen*) est la principale langue parlée en République d'Haïti, avec environ 11 millions de locuteurs. Elle est également parlée par la diaspora haïtienne, qui compte environ 2 millions de locuteurs, en particulier en Amérique du Nord. Parlé dans un contexte de diglossie, le *kreyòl* a encore des difficultés à atteindre la parité avec le français, qui est considéré comme une variété plus prestigieuse, malgré les efforts déployés en Haïti pour lui accorder un statut égal.

Dans cet article, nous présentons notre travail sur le développement de modèles auto-supervisés de traitement de la parole, entraînés exclusivement sur du *kreyòl*, [1] qui sont ensuite affinés sur une tâche de Reconnaissance Automatique de la Parole (RAP). Les modèles monolingues que nous entraînons représentent une rupture par rapport aux approches antérieures pour cette langue, qui s'appuyaient exclusivement sur de l'apprentissage par transfert à partir de modèles multilingues, ou à partir de modèles monolingues d'une langue bien dotée et typologiquement proche, et plus spécifiquement de sa langue lexificatrice, le français. Notre travail permet au *kreyòl* d'accéder à un statut égal à celui d'autres langues bien dotées telles que l'anglais (Baevski *et al.*, 2020, 2022), le français (Parcollet *et al.*, 2024) ou le chinois mandarin (Lu & Chen, 2023), qui disposent toutes de modèles monolingues dédiés. Dans notre travail, nous montrons notamment que nos modèles monolingues sont compétitifs par rapport à des modèles de taille similaire entraînés par apprentissage par transfert pour la reconnaissance automatique de la parole, bien qu'ayant moins de paramètres et ayant été entraînés sur moins de données.

Au-delà de l'aspect technique, notre travail a un poids symbolique important. Pendant trop longtemps, il a été affirmé que les langues créoles étaient destinées à disparaître au profit de leurs langues lexificatrice (un processus appelé « décréolisation », voir DeGraff (2003) pour un examen critique), en l'occurrence, le français. Ce processus a été théorisé sous la forme d'un cycle de décréolisation (cycle "pidgin → créole → décreolisation"). Notre travail montre que les créoles, et plus spécifiquement ici le *kreyòl ayisyen*, peuvent recevoir la même attention que les langues majeures et que les locuteurs de *kreyòl* devraient avoir et ont *désormais*, grâce à ce travail, un accès égal aux technologies de la parole pour leur langue, tout comme les locuteurs de langues bien dotées.

2 Contexte

Le domaine du traitement de la parole pour les langues créoles est relativement peu développé, comparé à d'autres langues telles que l'anglais ou le français, principalement en raison du manque de données accessibles. L'un des tout premiers corpus créés pour entraîner des modèles de traitement de la parole pour le *kreyòl*, et plus particulièrement pour entraîner des modèles de RAP, a été réalisé dans le cadre du projet DARPA DIPLOMAT à la fin des années 1990 (Frederking *et al.*, 1997), et est communément connu sous le nom de HAITI-CMU. Plus récemment, un plus grand corpus d'enregistrements a été collectée pour le programme IARPA BABEL, et est accessible par le biais du *Linguistic Data Consortium* (Andrus, Tony *et al.*, 2017) (voir section 3 pour plus de détails sur ces deux corpus). À notre connaissance, en dehors de ces deux corpus, il n'en existe pas

1. Les modèles sont disponibles sur https://huggingface.co/LLL-CREAM et la liste *exacte* et *exhaustive* des segments de parole utilisés est disponible sur https://gin.g-node.org/CREAM/SSL-Haitian permettant une reproductibilité totale.

(a)

Partition	Corpus Type	BABEL	CNHC	RADIO-H	ALH	H-CMU	VOXLG	**Total Marginal**
Train	Pré-VAD	176h 57m	5h 03m	1131h 11m	345h 36m	16h 04m	73h 47m	1748h 40m
	Post-VAD	74h 03m	4h 05m	1001h 15m	221h 55m	12h 18m	73h 47m	1387h 25m
Val	Pré-VAD	5h 25m	2h 03m	88h 08m	5h 16m	1h 29m	11h 32m	113h 57m
	Post-VAD	2h 11m	1h 35m	75h 03m	3h 32m	1h 07m	11h 32m	95h 02m
Test	Pré-VAD	20h 02m	2h 09m	114h 36m	5h 26m	2h 03m	10h 35m	154h 53m
	Post-VAD	8h 19m	1h 37m	100h 44m	4h 01m	1h 27m	10h 35m	126h 45m
Total	Pré-VAD	202h 26m	9h 16m	1333h 56m	356h 18m	19h 37m	95h 55m	2017h 31m
	Post-VAD	84h 34m	7h 17m	1177h 03m	229h 29m	14h 53m	95h 55m	**1609h 14m**

(b)

Corpus	Nombre	M	F	INC
ALH	216	117	42	57
BABEL	1091	547	544	–
CNHC	–	–	–	–
H-CMU	152	80	59	13
RADIO-H	1081	–	–	1081
VOXLG	7007	–	–	7007
Total Marginal	9547	744	645	8158

TABLE 1 – (1a) présente la distribution des durées d'enregistrement pour chaque partition. *Pré-VAD* fait référence à la durée des enregistrements avant l'application de tout algorithme de VAD, tandis que *Post-VAD* indique la durée des enregistrements après avoir conservé uniquement les segments identifiés comme de la parole par le VAD. (1b) présente la répartition en genre dans l'ensemble des corpus (M[asculin]/F[éminin]/INC[onnu]). Le symbole "–" est utilisé lorsque la valeur est inconnue. *Nombre* représente le nombre de locuteurs uniques. Chaque corpus est désigné par son acronyme, qui est introduit entre parenthèses après son nom complet dans section 3.

d'autre publiquement disponible avec des transcriptions standardisées, ce qui a sérieusement limité le développement d'outils pour cette langue. En réalité, cette pénurie de données ne se limite pas au domaine oral, mais également à l'écrit, et ce n'est que récemment que des initiatives ont cherché à développer des ressources textuelles ciblant spécifiquement les langues créoles.

La plupart des travaux sur le *kreyòl* se sont principalement concentrés sur la reconnaissance automatique de la parole (RAP) (Vu *et al.*, 2012; Breiter, 2014). Cependant, dans la plupart des travaux, la RAP est souvent traitée comme un composant d'un pipeline plus large de traduction parole vers texte (Frederking *et al.*, 1997, 2000), où les transcriptions servent d'entrée pour des modèles de traduction automatique dans une approche en cascade. Au-delà de la RAP, la recherche *utilisant* (plutôt que *sur*) le *kreyòl* a également exploré la détection de mots clés (Le *et al.*, 2014; Gales *et al.*, 2014), grâce à des corpus tels le *IARPA Babel Haitian Creole Language Pack* (Andrus, Tony *et al.*, 2017), qui a été spécifiquement créé à cette fin. La recherche ciblant d'autres créoles à base lexicale française est également relativement peu développé. Ce n'est que récemment que des modèles de RAP ont été développés pour le créole guadeloupéen, qui est apparenté au *kreyòl* et mutuellement intelligible (Macaire *et al.*, 2022; Le Ferrand *et al.*, 2023; Le Ferrand & Prud'hommeaux, 2024). En ce qui concerne le créole mauricien – parlé à Maurice, qui n'est ni apparenté au *kreyòl* ni mutuellement intelligible – le domaine semble plus actif, avec des contributions notables de (Noormamode *et al.*, 2019; Gooda Sahib-Kaudeer *et al.*, 2019; Gobin-Rahimbux *et al.*, 2023) et une partie du travail de (Macaire *et al.*, 2022), bien que la recherche reste également relativement limitée.

Malgré de récentes initiatives individuelles visant à créer des modèles modernes de traitement de la parole pour le *kreyòl* en utilisant l'apprentissage par transfert à partir de modèles multilingues pré-entraînés comme WHISPER (Radford *et al.*, 2023), XLS-R (Babu *et al.*, 2022) ou MMS (Pratap *et al.*, 2024a), des initiatives *reproductibles* et *scientifiques* pour développer des modèles monolingues comparables à WAV2VEC2 pour l'anglais ou LEBENCHMARK pour le français sont, à notre connaissance, inexistantes. La seule exception est le travail de Havard *et al.* (2024) qui est limité en termes de portée et de domaine, car il ne s'est appuyé que sur 250 heures de données orales, qui ne comportaient que des enregistrements de terrain, issus d'un unique corpus.

En résumé, il y a un écart important entre les modèles de traitement de la parole état de l'art disponibles

pour des langues comme l'anglais ou le français et ceux pour le *kreyòl*. Cet écart est largement dû au manque de données accessibles, qui empêche le développement de modèles d'envergure similaire. De plus, les précédents travaux sur le traitement de la parole pour le *kreyòl*, principalement des modèles de RAP, sont restés largement confidentiels et inaccessibles. Notre travail vise à relever ces défis en (1) donnant une vue d'ensemble des corpus oraux existants en *kreyòl ayisyen*, (2) entraînant des modèles monolingues auto-supervisés état de l'art tels que WAV2VEC2 et DATA2VEC, et (3) en affinant ces modèles sur une tâche de RAP tout en les mettant gratuitement et librement à la disposition de la communauté des créolistes, notre principal public cible.

3 Corpus de *kreyòl ayisyen*

Cette section présente une vue d'ensemble des corpus de parole en *kreyòl ayisyen* que nous utilisons dans nos expériences. La Table 1 offre une comparaison concise de chaque corpus, en soulignant leur taille, le partitionnement train/val/test et la répartition des locuteurs en genre.

ATLAS LINGUISTIQUE D'HAÏTI (ALH) L'ATLAS LINGUISTIQUE D'HAÏTI (Fattier, 1998) est une collection de 499 d'enregistrements audio collectés à Haïti entre 1978 et 1987 dans le but de créer un atlas linguistique, et totalise 356h d'audio. Les enregistrements ont été réalisés à l'origine sur des cassettes audio avec des magnétophones, puis numérisés par la Bibliothèque nationale de France (BNF) en 2010. Chaque enregistrement dure en moyenne 45 minutes et met en scène un ou plusieurs enquêteurs élicitant des mots ou des phrases auprès de leurs collaborateurs natifs. Ces données ont été rendues *publiques* et *légalement* accessibles par la *BNF* et la *Faculté de Linguistique Appliquée de l'Université d'État d'Haïti*, et sont téléchargeables *via* la plateforme COCOON (FLA & Fattier, 2015). Ce corpus est entièrement constitué de parole brute. Chaque enregistrement est enrichi de métadonnées détaillées, comprenant le lieu exact de la collecte et des informations sur le locuteur (notamment l'âge, la profession, le niveau d'alphabétisation). Les enregistrements couvrent l'ensemble du pays, garantissant une grande diversité d'accents et de vocabulaire, et contient plus de 200 de locuteurs.

CORPUS OF NORTHERN HAITIAN CREOLE (CNHC) Le CORPUS OF NORTHERN HAITIAN CREOLE (Valdman *et al.*, 2015) comporte 10 entretiens enregistrés et transcrits, menés au Cap-Haïtien (Nord d'Haïti) pour étudier les variations dialectales par rapport au haïtien standard, et totalise 9h d'audio. Les transcriptions utilisent une orthographe non standard pour représenter plus fidèlement les prononciations des locuteurs. Par exemple, les variations standard/non standard comprennent : "*Pòwoprens*/*Pòtoprens*" (Port-au-Prince, capitale d'Haïti) ; "*eske*/*eskeu*" (est-ce que) ; ou "*de*/*deu*" (deux). Bien que nous ne disposions pas du nombre exact de locuteurs, nous l'estimons entre 10 et 20, tous originaires de la région du Cap-Haïtien.

HAITI-CMU (H-CMU) Le corpus HAITI-CMU se compose d'enregistrements qui ont été initialement collectés pour le projet DARPA DIPLOMAT (Frederking *et al.*, 1997) à la fin des années 1990 et totalise 20h de parole. Le corpus contient uniquement de la parole lue, avec près de 7,5k phrases provenant de diverses sources (romans, discours, matériel religieux, etc.), prononcées par 152 locuteurs. Les enregistrements ont été collectés dans différents pays (notamment aux États-Unis & en France) avec des locuteurs de différentes origines et régions d'Haïti. Certains des locuteurs ayant vécu ou ayant été élevés en France métropolitaine ou dans des Territoires d'Outre-Mer où des créoles voisins (guadeloupéen, martiniquais, guyanais) sont parlés, cela a pu avoir une incidence sur leur accent. Tous les enregistrements sont transcrits en utilisant les conventions orthographiques standard

du *kreyòl*. Les enregistrements ont été faits dans un environnement calme permettant d'obtenir ainsi des enregistrements de grande qualité.

IARPA-BABEL HAITIAN CREOLE (BABEL) Le corpus IARPA-BABEL HAITIAN CREOLE (Andrus, Tony *et al.*, 2017) est constitué d'enregistrements en *kreyòl* qui ont été enregistré dans le cadre du programme BABEL. Les enregistrements ont été collectés en 2012/2013 et se composent principalement de conversations téléphoniques. Certains enregistrements ont également été collectés à l'aide de matériel scripté, où il a été demandé aux locuteurs de lire des phrases spécifiques et/ou de discuter de sujets spécifiques. Le corpus se compose de plus de 11k enregistrements, totalisant plus de 202 heures de parole pour plus de 1000 locuteurs. La plupart des enregistrements ont été transcrits en utilisant les conventions orthographiques standard, mais certains n'ont pas été transcrits. Afin de garantir une diversité maximale, les locuteurs proviennent de différentes régions d'Haïti, et couvrent une large gamme d'âge, et sont répartis de manière équilibrée entre hommes et femmes. Les enregistrements ont été effectués dans des conditions environnementales variées, allant d'environnements très bruyants (par exemple dans la rue) à des environnements très calmes (par exemple, à la maison ou au bureau).

VOXLINGUA HAITIAN (VOXLG) VoxLingua (Valk & Alumäe, 2021) est corpus qui comporte uniquement des enregistrements audio récupérés sur YouTube, afin d'entraîner des modèles d'identification de langue. Le corpus contient des enregistrements pour 107 langues. Dans ce travail, nous n'utilisons que les enregistrements qui ont été identifés comme étant en *kreyòl*, soit un total de 96h de parole. Les enregistrements de ce corpus ont déjà été pré-traités (Valk & Alumäe, 2021) pour séparer les sections contenant de la parole du bruit de fond et des segments non parlés à l'aide d'un modèle de détection de l'activité vocale (VAD). Un modèle de segmentation en locuteurs (*diarisation*) a également été appliqué pour séparer les locuteurs les uns des autres, ce qui a permis d'obtenir plus de $36,7$k segments pour 664 enregistrements à l'origine. Selon le modèle de diarisation, le corpus se compose de plus de 7k locuteurs uniques. Étant donné qu'il n'y a pas d'annotations manuelles pour ce corpus, il ne s'agit que d'une estimation qui doit être interprétée avec prudence, car ce chiffre semble largement surestimer le nombre réel de locuteurs.

RADIO HAITI (RADIO HAITI) Alors que les corpus précédemment mentionnés sont connus de la communauté du traitement de la parole, et ont déjà été utilisés dans d'autres publications, RADIO HAITI est un nouveau corpus, qui est utilisé pour la première fois dans cette publication. RADIO HAITI se compose d'enregistrements diffusés par *Radio Haïti-Inter* entre 1957 et 2003, qui ont été mis à disposition par l'Université de Duke par l'intermédiaire de la *David M. Rubenstein Rare Book Manuscript Library* (Duke University & Various Contributors, 2024). Le corpus contient plus de 5000 enregistrements totalisant 2400 heures de contenu. Environ la moitié des enregistrements sont uniquement constitués de parole en *kreyòl*, tandis que l'autre moitié est principalement en français ou dans un mélange de français et de *kreyòl*, et, dans une moindre mesure, en anglais ou en espagnol. Dans ce travail, nous n'utilisons qu'un sous-ensemble de ce corpus, en particulier la partie en *kreyòl*. Cette partie se compose d'environ $2,9$k enregistrements totalisant plus de $1,3$k heures. Il s'agit, à notre connaissance, de la plus large collection d'enregistrements en *kreyòl*.

Bien que la plupart des enregistrements aient été réalisés en studio, garantissant ainsi un son de haute qualité et exempt de bruit, le corpus comprend également des interviews et des interventions réalisées par téléphone ou dans des espaces publics, ce qui introduit un large éventail de conditions acoustiques. Chaque enregistrement est associé à des métadonnées qui précisent les principaux sujets abordés et les principaux intervenants, généralement les présentateurs et, le cas échéant, les personnes interrogées. Bien qu'il soit impossible de déterminer le nombre exact de locuteurs (par exemple, les

locuteurs qui interviennent au téléphone ou pendant les entretiens ne sont pas mentionnés dans les métadonnées), les enregistrements que nous utilisons comportent *au moins* 1,000 individus uniques. Cependant, même si nous avons remarqué que les locuteurs sont d'origines différentes et ont, pour certains, un accent distinct (par exemple, de la région du Cap Haïtien), nous ne disposons pas de métadonnées indiquant l'origine des locuteurs. Enfin, les métadonnées précisent la date de diffusion des enregistrements (au moins l'année). La plupart des enregistrements utilisés dans cette publication ont été diffusés entre les années 1980 et le début des années 2000.

Dans l'ensemble, le travail que nous présentons dans cet article utilise plus de 2000h d'enregistrements en *kreyòl ayisyen*. Bien que nous ne puissions pas donner le nombre exact de locuteurs unique, nous l'estimons à plus de 3000.[2] Les enregistrements couvrent un large éventail de conditions environnementales, allant d'enregistrements de terrain (ALH) à des enregistrements réalisés dans des environnements calmes (CNHC), avec de la parole lue (H-CMU, BABEL) ou des conversations plus naturelles (BABEL, VOXLG, RADIO-H). En ce qui concerne l'équilibre en genre, bien que nous ne disposions pas de cette information pour tous les corpus que nous utilisons, lorsque cette information est disponible, nous observons une répartition assez équilibrée avec 645 femmes et 744 hommes (voir Table 1b).

4 Pré-entraînement

Pour assurer une cohérence entre tous les corpus, nous avons appliqué un pipeline de prétraitement standardisé à chaque corpus avant d'entraîner nos modèles auto-supervisés de traitement de la parole. Tout d'abord, tous les enregistrements ont été rééchantillonnés en PCM mono 16 bits, 16 kHz. Un modèle de détection de l'activité vocale (VAD) (Bredin *et al.*, 2020; Bredin & Laurent, 2021) a ensuite été appliqué à chaque enregistrement afin d'isoler les sections parlées des silences et de la non-parole. Les segments obtenus ont ensuite été combinés en les concaténant de manière itérative, en préservant l'ordre temporel afin d'atteindre une durée moyenne de 20 secondes.[3] Dans les rares cas où les segments VAD étaient plus grands que 30s, nous avons procédé à un rognage dur en les coupant en sous-segments d'une longueur maximale de 30s.[4] Les segments concaténés ont atteint une durée moyenne de 21.9s avec un écart type de 7.3s. Étant donné que VOXLG a déjà été pré-traité de manière similaire par (Valk & Alumäe, 2021), nous n'avons pas

Model	Dim. Sortie	Dim. Inter.	Tête Attn.	Blocks Transf.	Etapes Entraîn.	# GPUs Entraîn.	Heures GPU
data2vec-base	768	3072	12	12	292k	16	3626.65
wav2vec2-base	768	3072	12	12	225k	8	3739.67
wav2vec2-large	1024	4096	16	24	218k	16	4340.99

TABLE 2 – Hyperparamètres des meilleurs modèles-étape (*checkpoints*), avec notamment les dimensions (interne et de sortie) des blocs de Transformer, le nombre de têtes d'attention et les statistiques d'entraînement (dont le nombre d'étapes (*steps*), le nombre de GPUs utilisés, et le nombre d'heures-GPU).

2. Compte tenu des surestimations de VOXLINGUA HAITIAN estimées par le modèle de diarisation utilisé par (Valk & Alumäe, 2021), il nous semble plus raisonnable de ne prendre en compte que 1,000 locuteurs.

3. Cette valeur a été choisie de manière à obtenir des segments de longueur similaire en tenant compte de l'écart-type tout en limitant le nombre de segments d'une longueur supérieure à 30s. Pour référence, (Baevski *et al.*, 2020) et (Parcollet *et al.*, 2024) utilisent des segments de 20 secondes pour les modèles BASE et LARGE, tandis que (Parcollet *et al.*, 2024) utilise des segments de 15 secondes pour leur modèle XLARGE, et (Conneau *et al.*, 2021) utilise des segments de 15 secondes et 20 secondes pour leurs modèles BASE et LARGE, respectivement.

4. Ceci pourrait potentiellement aboutir à la division d'un même mot sur deux segments, cependant, le nombre de coupes ainsi effectué est minime, et cette méthode n'a été appliquée qu'aux segments du corpus RADIO-H.

Model	Global				Genre								Corpus							
	CER		WER		CER				WER				CER				WER			
					F		M		F		M		BABEL		H-CMU		BABEL		H-CMU	
	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std
data2vec-hat-1.4K-base	**21.0**	51.4	**33.7**	49.9	**19.3**	41.0	**23.0**	60.2	**31.9**	43.4	**35.7**	55.8	**22.9**	53.4	**1.5**	6.1	**36.4**	51.3	5.0	9.5
wav2vec2-hat-1.4K-large	21.3	53.0	34.8	49.7	19.7	43.2	23.1	61.5	33.2	43.9	36.8	55.1	23.1	55.1	1.4	6.1	37.6	51.1	5.3	9.7
wav2vec2-hat-1.4K-base	21.5	52.0	34.5	52.8	19.7	41.1	23.5	61.1	32.7	44.1	36.8	60.4	23.4	54.0	**1.5**	6.4	37.3	54.3	**4.8**	9.1
xlsr2-300m-large	21.5	50.1	35.5	51.0	19.7	39.4	23.6	59.1	33.6	43.4	37.7	57.7	23.4	52.0	1.6	5.9	38.3	52.4	5.9	10.1
wav2vec2-FR-7K-large	22.5	51.7	36.8	48.6	20.8	42.0	24.5	60.0	35.3	43.7	38.7	53.1	24.5	53.6	**1.5**	6.0	39.8	49.7	5.3	9.7
wav2vec2-FR-3K-large	23.4	51.5	38.3	48.5	21.5	41.2	25.5	60.2	36.6	42.3	40.4	54.1	25.4	53.4	1.6	6.1	41.4	49.6	5.4	9.6
wav2vec2-FR-7K-base	23.8	52.9	38.4	50.6	22.0	42.7	25.8	61.5	37.0	45.1	40.2	55.7	25.9	54.8	1.6	6.2	41.5	51.8	5.5	9.6
wav2vec2-FR-2.6K-base	24.6	54.8	39.4	53.9	22.9	46.7	26.5	62.0	37.6	47.1	41.6	60.1	26.7	56.8	1.6	6.3	42.6	55.3	5.3	9.5
wav2vec2-FR-3K-base	24.8	55.5	40.0	54.7	23.4	47.1	26.5	63.1	38.1	45.9	42.3	62.4	27.0	57.6	1.8	6.4	43.2	56.1	6.2	9.9
wav2vec2-FR-1K-large	26.5	48.7	45.8	47.8	24.7	40.0	28.6	56.1	43.8	42.9	48.4	52.3	28.8	50.3	2.2	6.5	49.3	48.4	8.4	11.1
wav2vec2-FR-1K-base	27.0	54.8	44.2	53.4	25.1	47.3	29.1	61.6	41.9	46.0	47.0	59.9	29.3	56.7	1.9	6.0	47.7	54.4	6.9	13.1
mms-1b-all	30.7	51.3	58.2	49.0	29.3	45.5	32.3	56.6	56.4	43.8	60.4	53.8	33.2	52.9	3.6	6.9	62.1	49.3	16.1	14.9
mms-1b-fl102	31.6	49.3	61.7	46.9	30.2	44.8	33.4	53.5	59.8	43.3	64.0	50.3	34.1	50.8	5.3	6.9	65.2	47.3	24.2	15.8
xlsr53-56k	34.7	54.2	66.4	50.7	33.2	46.9	36.7	60.8	64.2	46.2	69.2	54.9	37.5	55.9	5.6	7.0	70.2	51.2	26.0	16.5

TABLE 3 – Taux d'erreur mot (WER) et caractère (CER) moyen ($\pm$ écart-type) au niveau global, par genre et par corpus, triés par CER global sur l'ensemble de test (meilleur modèle sélectionné sur la base des résultats de validation). Le CER le plus bas est en **bleu**, le WER le plus bas est en **rouge**, et le WER/CER moyen le plus bas par catégorie (F/M ou BABEL/H-CMU) est en gras. Les noms de modèles en gras indiquent des versions affinées de nos modèles pré-entraînés.

appliqué notre pipeline à ce corpus. Dans l'ensemble, l'utilisation d'une étape de prétraitement VAD a entraîné une réduction de 25% de la quantité totale d'audio. Cette valeur varie considérablement d'un ensemble de données à l'autre, allant de 57% pour BABEL à seulement 11% pour RADIO-H (voir Table 1a).

Les partitions train/val/test ont été construites de manière à garantir que l'ensemble de validation inclut des locuteurs non vus dans la partition train et que l'ensemble de test inclut des locuteurs non vus dans les partitions train et val. Cela permet de garantir la sélection d'un modèle qui généralise mieux en termes de locuteurs.

Nous entraînons trois type de modèles auto-supervisés : WAV2VEC2-BASE, WAV2VEC2-LARGE et DATA2VEC-AUDIO-BASE. [5] Pour assurer une reproductibilité maximale et en cohérence avec les travaux antérieurs, tous nos modèles ont été entraînés à l'aide de PyTorch (Paszke *et al.*, 2019) par le biais de la boîte à outils `fairseq` (Ott *et al.*, 2019). Les hyperparamètres de l'architecture des modèles WAV2VEC2-BASE et WAV2VEC2-LARGE sont identiques à (Evain *et al.*, 2021), et ceux du modèle DATA2VEC sont identiques à (Baevski *et al.*, 2022). Les architectures ne diffèrent que modérément de celles présentées dans (Baevski *et al.*, 2020) avec par exemple 12 de têtes d'attention pour notre version BASE au lieu de 8. Au lieu d'entraîner les modèles pour un nombre fixe d'étapes (par exemple, (Parcollet *et al.*, 2024) entraîne ses modèles BASE et LARGE de 1k pour 200k étapes), nous avons décidé d'arrêter manuellement l'entraînement lorsque les modèles avaient convergé, la convergence étant définie comme le point où les courbes d'entraînement et de validation se croisent. Les principaux hyperparamètres de nos modèles sont indiqués dans Table 2.

5. DATA2VEC-AUDIO-BASE utilise un squelette d'architecture WAV2VEC2-BASE.

5 Affinage

5.1 Paramètres expérimentaux

Une fois nos modèles WAV2VEC2-BASE, WAV2VEC2-LARGE et DATA2VEC-AUDIO-BASE pré-entraînés, nous les avons affinés sur une tâche de reconnaissance de la parole (RAP). En plus de nos trois modèles pré-entraînés, nous avons également affiné `XLSR53-56k` et `XLSR2-300m-large` (Babu *et al.*, 2022, pour chacun, l'ensemble du modèle a été affiné) et `MMS-1B-all` et `MMS-1B-FL102` (Pratap *et al.*, 2024b, en affinant seulement des couches d'adaptation). Pour ce faire, nous avons utilisé les enregistrements transcrits des corpus BABEL et H-CMU. Contrairement à CNHC, les deux comportent des transcriptions qui suivent l'orthographe standard conçue par l'Institut Pédagogique National (Valdman, 1982), qui est couramment utilisée par les institutions officielles haïtiennes.

Nous avons réutilisé les mêmes partitions de train/val/test que celles utilisées pour le pré-entraînement, à la seule différence que nous avons exclu les enregistrements du corpus BABEL qui n'étaient pas transcrits. Par conséquent, les partitions de train/val/test utilisées pour l'affinage sont des sous-ensembles des partitions de pré-entraînement. Pour les deux corpus, les locuteurs de la partition de validation et ceux de la partition de test ne sont pas présents dans la partition d'entraînement. De plus, les locuteurs de la partition de test ne figurent pas non plus dans la partition de validation. Contrairement au pré-entraînement, où nous avons utilisé des segments concaténés obtenus par un modèle VAD, nous avons utilisé ici la segmentation de référence pour segmenter les enregistrements en segments plus courts. Les modèles ont été entraînés avec une fonction de coût CTC et gelés les 10k premiers pas pour éviter un sur-entraînement. Nous avons utilisé le décodage Viterbi sans aucun modèle de langue externe, avec un faisceau (*beam search*) de 5.

5.2 Pré-traitement du texte

Nous utilisons un vocabulaire de 28 symboles (27 caractères et un espace supplémentaire ⟨|⟩).[6] Lorsque nous avons rencontré des graphèmes étrangers dans des mots d'emprunt ou des noms propres (par exemple ⟨qu⟩ dans "Inuqua"), nous les avons convertis en leur équivalent le plus proche (par exemple, ⟨k⟩ comme dans "Inuka"). Les chiffres ont été intégralement épelés (par exemple "21" : "venteyen") à l'aide du package `num2words` que nous avons adapté pour le *kreyòl*,[7] et les abréviations ont été intégralement épelées (par exemple "HTML" : "ach te èm èl") à l'aide d'un script personnalisé. Toutes les annotations spécifiques au corpus BABEL (par exemple, pour signaler le bruit, les hésitations, etc.) ont été supprimées. Enfin, les transcriptions ont été mises en minuscules, toute la ponctuation a été supprimée et les caractères ont été normalisés à l'aide de la norme NFKC Unicode (Unicode, 2024). Enfin, nous avons séparé les pronoms clitiques ("t", "m", "n", "y", "l", et "w", comme par exemple "w" comme dans "papa w", *votre père*) des mots auxquels ils se rattachent. Le *kreyòl* étant avant tout une langue orale, leur écriture varie considérablement, utilisant souvent des traits d'union, des apostrophes, des espaces ou des concaténations directes sans régularité stricte.

6. `aàbcdeèfghijklmnoòprstuvwyz|`
7. Disponible sur https://github.com/LLL-Orleans/num2words

5.3 Résultats & Discussion

La Table 3 montre les taux d'erreurs mots et caractères (WER/CER) globaux, puis par sexe et par corpus. Les meilleurs modèles sont nos modèles monolingues pré-entraînés sur $1,4$K heures de parole en *kreyòl* (avec un CER de 21 et un WER de 33.7 pour le meilleur d'entre eux, DATA2VEC-AUDIO-BASE), suivis de près par `XLSR-300m-large`. De manière surprenante, bien qu'ils aient été affinés sur les mêmes données, `XLSR53-56k` et les deux modèles `MMS` se placent loin derrière. Nous observons que nos modèles, bien qu'ayant moins de paramètres (95M pour les modèles BASE et 317M pour LARGE) ont obtenu de meilleurs scores que des modèles plus grands (`MMS-1B`), et de meilleurs scores que des modèles entraînés sur beaucoup plus d'heures d'enregistrement et plus de langues (1k langues pour `MMS`, 100+ pour `XLSR`). Il est surprenant de constater que les modèles affinés à partir des modèles WAV2VEC2 français soient inférieurs à `XLSR-300m-large`, alors que le français est très proche du *kreyòl*. De précédents travaux (Lehečka *et al.*, 2023) ont montré qu'un transfert ciblé depuis une langue proche (par exemple, du tchèque vers le slovaque) peut surpasser les modèles multilingues de grande taille comme `XLSR`. Dans notre cas, même nos transferts les plus ciblés — des affinages à partir de modèles WAV2VEC2 français — sont moins performants que `XLSR-300m-large`, alors que le français est pourtant considéré comme proche du HC.

Nous observons des écarts types très élevés pour tous les modèles. Nous attribuons ce phénomène aux différences substantielles de taux d'erreur entre H-CMU et BABEL. H-CMU contient essentiellement de la parole lue, enregistrée dans des environnements calmes, tandis que BABEL contient majoritairement de la parole naturelle enregistrée dans des environnements bruyants à 8kHz sur des téléphones. La différence entre ces deux corpus réside à la fois dans la *nature* du contenu (parole lu vs. parole naturelle) et dans la qualité acoustique intrinsèque des enregistrements, BABEL étant connu pour être un *"benchmark de reconnaissance vocale difficile"*. (Babu *et al.*, 2022).

Enfin, nous avons étudié le taux d'erreur en fonction du sexe. Comme cela a été largement démontré (Adda-Decker & Lamel, 2005), le taux d'erreur est plus faible pour les femmes que pour les hommes. La différence moyenne entre les hommes et les femmes sur l'ensemble des modèles est de 4 points WER et de $3,6$ points de CER, avec de grandes variations entre les modèles (5 points de WER pour le modèle ajusté à partir de `wav2vec2-FR-1K-base` et $3,2$ points pour `wav2vec2-FR-7K`). Nos modèles monolingues se situent entre les deux extrêmes, avec le modèle affiné à partir de `wav2vec2-hat-1.4K-base` ayant une différence de 4 en points de WER entre les hommes et les femmes.

En résumé, nos modèles monolingues *kreyòl*, pré-entraînés sur 1700 heures de parole brute (voir Table 1a), sont plus performants que les modèles multilingues plus importants, ce qui souligne la valeur du pré-entraînement spécifique au domaine.

6 Conclusion

Ainsi dans ce travail, nous avons présenté trois modèles pré-entraînés de traitement de la parole pour le *kreyòl* basés sur les architectures WAV2VEC2-BASE, WAV2VEC2-LARGE, et DATA2VEC-AUDIO-BASE. Nous avons affiné ces modèles sur une tâche de reconnaissance de la parole et montré qu'ils obtenaient des résultats compétitifs par rapport à des modèles multilingues et monolingues de plus grande échelle. C'est la première fois que des modèles de cette envergure sont développés pour le *kreyòl*, l'établissant désormais comme une langue bien dotée en termes de modèles auto-supervisés

de traitement de la parole.

Notre travail fournit ainsi les outils nécessaires pour une reconnaissance robuste de la parole, et des modèles sources pour l'apprentissage par transfert vers d'autres créoles caribéens basés sur le français, et contribue ainsi à accorder au *kreyòl* un statut égal à celui du français. Nous espérons que les créolistes s'empareront de ces modèles pour transcrire leur données à grande échelle. Nos modèles ont déjà été utilisés pour transcrire 1400 heures d'enregistrements de RADIO HAITI, le tout en moins de 2 heures, permettant une étude sans précédent des phénomènes linguistiques du *kreyòl*.

Remerciements

Cette recherche a été financée par l'Agence nationale de la recherche (ANR) au titre du projet ANR-20-CE38-0006 (projet CREAM). Les expériences ont été menées à l'aide de Grid'5000, développé sous INRIA ALADDIN avec l'appui du CNRS, RENATER et diverses universités ; de CaSciModOT au Centre de Calcul Scientifique en région Centre-Val de Loire ; et les ressources HPC de IDRIS fournies par GENCI (allocation 2024-AD011014940).

Références

ADDA-DECKER M. & LAMEL L. (2005). Do speech recognizers prefer female speakers ? In *Interspeech 2005*, p. 2205–2208. DOI : 10.21437/Interspeech.2005-699.

ANDRUS, TONY, BILLS, ARIC, CONNERS, THOMAS, CRABB, ERIN SMITH, DUBINSKI, EYAL, FISCUS, JONATHAN G., GILLIES, BREANNA, HARPER, MARY, HAZEN, T. J., HEFRIGHT, BROOK, JARRETT, AMY, LE, HANH, RAY, JESSICA, RYTTING, ANTON, SHEN, WADE, SILBER, RONNIE, TZOUKERMANN, EVELYNE & BISHOP J. (2017). IARPA Babel Haitian Creole Language Pack IARPA-babel201b-v0.2b. DOI : 10.35111/ENHB-6110.

BABU A., WANG C., TJANDRA A., LAKHOTIA K., XU Q., GOYAL N., SINGH K., VON PLATEN P., SARAF Y., PINO J., BAEVSKI A., CONNEAU A. & AULI M. (2022). XLS-R : Self-supervised cross-lingual speech representation learning at scale. In *Interspeech 2022*, p. 2278–2282. DOI : 10.21437/Interspeech.2022-143.

BAEVSKI A., HSU W., XU Q., BABU A., GU J. & AULI M. (2022). data2vec : A general framework for self-supervised learning in speech, vision and language. In K. CHAUDHURI, S. JEGELKA, L. SONG, C. SZEPESVÁRI, G. NIU & S. SABATO, Éds., *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 de *Proceedings of Machine Learning Research*, p. 1298–1312 : PMLR.

BAEVSKI A., ZHOU H., MOHAMED A. & AULI M. (2020). wav2vec 2.0 : a framework for self-supervised learning of speech representations. In *NeurIPS*, NIPS '20, Red Hook, NY, USA : Curran Associates Inc.

BREDIN H. & LAURENT A. (2021). End-to-end speaker segmentation for overlap-aware resegmentation. In *Proc. Interspeech 2021*, Brno, Czech Republic.

BREDIN H., YIN R., CORIA J. M., GELLY G., KORSHUNOV P., LAVECHIN M., FUSTES D., TITEUX H., BOUAZIZ W. & GILL M.-P. (2020). pyannote.audio : neural building blocks for speaker diarization. In *ICASSP 2020*, Barcelona, Spain.

BREITER W. (2014). Rapid bootstrapping of haitian creole large vocabulary continuous speech recognition.

CONNEAU A., BAEVSKI A., COLLOBERT R., MOHAMED A. & AULI M. (2021). Unsupervised cross-lingual representation learning for speech recognition. In *Interspeech 2021*, p. 2426–2430. DOI : 10.21437/Interspeech.2021-329.

DEGRAFF M. (2003). Against creole exceptionalism. *Language*, **79**(2), 391–410.

DUKE UNIVERSITY & VARIOUS CONTRIBUTORS (2024). Radio haiti collection. https://idn.duke.edu/ark:/87924/m1j07w. Digital items : 5,314 ; Total components : 3660 ; Last Indexed : 2024-12-05.

EVAIN S., NGUYEN M. H., LE H., ZANON BOITO M., MDHAFFAR S., ALISAMIR S., TONG Z., TOMASHENKO N., DINARELLI M., PARCOLLET T., ALLAUZEN A., ESTÈVE Y., LECOUTEUX B., PORTET F., ROSSATO S., RINGEVAL F., SCHWAB D. & BESACIER L. (2021). Task Agnostic and Task Specific Self-Supervised Learning from Speech with LeBenchmark. In *Thirty-fifth Conference on NeurIPS (NeurIPS 2021)*, NeurIPS 2021 Datasets and Benchmarks Track, on-line, United States. HAL : hal-03407172.

FATTIER D. (1998). *Contribution à l'étude de la genèse d'un créole : l'Atlas linguistique d'Haïti, cartes et commentaires, 6 vol.* Bibliographical record, Presses Universitaires du Septentrion, Villeneuve d'Ascq. Ph.D. Dissertation, Université de Provence.

FLA F. & FATTIER D. (2015). Atlas linguistique d'Haïti. DOI : 10.34847/COCOON.8EA988D2-BF16-303D-81A0-0C55CC035240.

FREDERKING R., RUDNICKY A. & HOGAN C. (1997). Interactive speech translation in the DIPLOMAT project. In *Spoken Language Translation*.

FREDERKING R., RUDNICKY A., HOGAN C. & LENZO K. (2000). Interactive speech translation in the diplomat project. *Machine Translation*, **15**(1/2), 27–42. DOI : 10.1023/A :1011172330853.

GALES M. J. F., KNILL K., RAGNI A. & RATH S. P. (2014). Speech recognition and keyword spotting for low-resource languages : Babel project research at cued. In *SLTU*.

GOBIN-RAHIMBUX B., GOODA SAHIB N., PEERTHY N., TAYLOR A. & ABDELLAH I. (2023). A voice-based personal assistant for mental health in kreol morisien. *JECE*, **2023**. DOI : 10.1155/2023/5532967.

GOODA SAHIB-KAUDEER N., GOBIN-RAHIMBUX B., BAHSU B. S. & MAGHOO M. F. A. (2019). Automatic speech recognition for kreol morisien : A case study for the health domain. In A. A. SALAH, A. KARPOV & R. POTAPOVA, Éds., *Speech and Computer*, p. 414–422, Cham : Springer International Publishing.

HAVARD W. N., GOVAIN R., GONÇALVES TEIXEIRA D., LECOUTEUX B. & SCHANG E. (2024). Technologies de la parole et données de terrain : le cas du créole haïtien. In M. BALAGUER, N. BENDAHMAN, L.-M. HO-DAC, J. MAUCLAIR, J. G MORENO & J. PINQUIER, Éds., *Actes de la 31ème Conférence sur le Traitement Automatique des Langues Naturelles, volume 1 : articles longs et prises de position*, p. 686–694, Toulouse, France : ATALA and AFPC.

LE V.-B., LAMEL L., MESSAOUDI A., HARTMANN W., GAUVAIN J.-L., WOEHRLING C., DESPRES J. & ROY A. (2014). Developing STT and KWS systems using limited language resources. In *Interspeech 2014*, p. 2484–2488. DOI : 10.21437/Interspeech.2014-527.

LE FERRAND E., PIERRE-LOUIS C., DONG R., LECOUTEUX B., GONÇALVES-TEIXEIRA D., HAVARD W. N. & SCHANG E. (2023). Outiller la documentation des langues créoles. In *LIFT 2023 : journées scientifiques du GdR Linguistique Informatique, Formelle et de Terrain*, Vandoeuvre-Lès-Nancy, France. HAL : hal-04302623.

LE FERRAND É. & PRUD'HOMMEAUX E. (2024). Automatic transcription of grammaticality judgements for language documentation. In S. MOELLER, G. AGYAPONG, A. ARPPE, A. CHAUDHARY, S. RIJHWANI, C. COX, R. HENKE, A. PALMER, D. ROSENBLUM & L. SCHWARTZ, Éds., *COMPUTEL*, p. 33–38, St. Julians, Malta : Association for Computational Linguistics.

LEHEČKA J., PSUTKA J. V. & PSUTKA J. (2023). Transfer learning of transformer-based speech recognition models from czech to slovak. In K. EKŠTEIN, F. PÁRTL & M. KONOPÍK, Éds., *Text, Speech, and Dialogue*, p. 328–338, Cham : Springer Nature Switzerland.

LU K.-H. & CHEN K.-Y. (2023). A context-aware knowledge transferring strategy for ctc-based asr. In *SLT2022*, p. 60–67. DOI : 10.1109/SLT54892.2023.10022825.

MACAIRE C., SCHWAB D., LECOUTEUX B. & SCHANG E. (2022). Automatic speech recognition and query by example for creole languages documentation. In S. MURESAN, P. NAKOV & A. VILLAVICENCIO, Éds., *ACL 2022*, p. 2512–2520, Dublin, Ireland.

NOORMAMODE W., GOBIN-RAHIMBUX B. & PEERBOCCUS M. (2019). A speech engine for mauritian creole. In S. C. SATAPATHY, V. BHATEJA, R. SOMANAH, X.-S. YANG & R. SENKERIK, Éds., *Information Systems Design and Intelligent Applications*, p. 389–398, Singapore : Springer Singapore.

OTT M., EDUNOV S., BAEVSKI A., FAN A., GROSS S., NG N., GRANGIER D. & AULI M. (2019). fairseq : A fast, extensible toolkit for sequence modeling. p. 48–53, Minneapolis, Minnesota : Association for Computational Linguistics. DOI : 10.18653/v1/N19-4009.

PARCOLLET T., NGUYEN H., EVAIN S., ZANON BOITO M., PUPIER A., MDHAFFAR S., LE H., ALISAMIR S., TOMASHENKO N., DINARELLI M., ZHANG S., ALLAUZEN A., COAVOUX M., ESTÈVE Y., ROUVIER M., GOULIAN J., LECOUTEUX B., PORTET F., ROSSATO S., RINGEVAL F., SCHWAB D. & BESACIER L. (2024). Lebenchmark 2.0 : A standardized, replicable and enhanced framework for self-supervised representations of french speech. *Computer Speech & Language*, **86**, 101622. DOI : https ://doi.org/10.1016/j.csl.2024.101622.

PASZKE A., GROSS S., MASSA F., LERER A., BRADBURY J., CHANAN G., KILLEEN T., LIN Z., GIMELSHEIN N., ANTIGA L., DESMAISON A., KÖPF A., YANG E., DEVITO Z., RAISON M., TEJANI A., CHILAMKURTHY S., STEINER B., FANG L., BAI J. & CHINTALA S. (2019). *PyTorch : an imperative style, high-performance deep learning library*, In *NeurIPS*. Curran Associates Inc. : Red Hook, NY, USA.

PRATAP V., TJANDRA A., SHI B., TOMASELLO P., BABU A., KUNDU S., ELKAHKY A., NI Z., VYAS A., FAZEL-ZARANDI M. *et al.* (2024a). Scaling speech technology to 1,000+ languages. *Journal of Machine Learning Research*, **25**(97), 1–52.

PRATAP V., TJANDRA A., SHI B., TOMASELLO P., BABU A., KUNDU S., ELKAHKY A., NI Z., VYAS A., FAZEL-ZARANDI M., BAEVSKI A., ADI Y., ZHANG X., HSU W.-N., CONNEAU A. & AULI M. (2024b). Scaling speech technology to 1,000+ languages. *Journal of Machine Learning Research*, **25**(97), 1–52.

RADFORD A., KIM J. W., XU T., BROCKMAN G., MCLEAVEY C. & SUTSKEVER I. (2023). Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23 : JMLR.org.

UNICODE (2024). UAX #15 : Unicode Normalization Forms.

VALDMAN A. (1982). *Education Reform and the Instrumentalization of the Vernacular in Haiti*, In B. HARTFORD, A. VALDMAN & C. R. FOSTER, Éds., *Issues in International Bilingual Education : The Role of the Vernacular*, p. 139–170. Springer US : Boston, MA. DOI : 10.1007/978-1-4684-4235-9_7.

VALDMAN A., VILLENEUVE A.-J. & SIEGEL J. F. (2015). On the influence of the standard norm of haitian creole on the cap haïtien dialect : Evidence from sociolinguistic variation in the third person singular pronoun. *Journal of Pidgin and Creole Languages*, **30**(1), 1–43. DOI : 10.1075/jpcl.30.1.01val.

VALK J. & ALUMÄE T. (2021). VoxLingua107 : a dataset for spoken language recognition. In *Proc. IEEE SLT Workshop*.

VU N. T., BREITER W., METZE F. & SCHULTZ T. (2012). Initialization schemes for multilayer perceptron training and their impact on asr performance using multilingual data. In *Interspeech 2012*, p. 2586–2589. DOI : 10.21437/Interspeech.2012-12.