Actes de CORIA-TALN 2026
Nantes, France, 29 juin au 3 juillet 2026

Apprentissage actif pour l'annotation morphosyntaxique du créole haïtien

Rayan Ziane[1,3] Maximin Coavoux[2] Benjamin Lecouteux[2] Emmanuel Schang[1]
(1) LLL, Université d'Orléans - BnF, F-45000 Orléans, France
(2) LIG, Université Grenoble Alpes, CNRS, Grenoble INP, F-38000 Grenoble, France
(3) CRISCO, Université de Caen, F-14000 Caen, France
<first>.<last>@{univ-orleans.fr, univ-grenoble-alpes.fr}

RÉSUMÉ

Cet article présente une méthodologie pour l'étiquetage morphosyntaxique des transcriptions du corpus radiophonique RADIO HAÏTI-INTER (1 300 heures) en créole haïtien parlé. Face au manque de données annotées pour l'oral, nous procédons d'abord à l'adaptation d'un modèle multilingue (XLM-RoBERTa) par pré-entraînement continué sur le corpus cible, puis à un premier affinage sur les treebanks Universal Dependencies existants. Nous évaluons une stratégie d'apprentissage actif guidée par une sélection des échantillons d'entraînement selon les scores de confiance du modèle (aléatoire, faible confiance, haute confiance) et deux stratégies d'affinage (séquentiel et joint). Les résultats montrent que l'adaptation au domaine est cruciale (gains de +4,3 points), que l'approche séquentielle surpasse l'affinage joint, mais que la sélection active n'apporte pas d'avantage significatif par rapport à un échantillonnage aléatoire. Nous mettons à disposition un échantillon annoté manuellement, un modèle de langue adapté au haïtien et un modèle d'étiquetage POS pour le haïtien parlé transcrit.

ABSTRACT

Active learning for part-of-speech tagging of Haitian Creole

This paper presents a methodology for part-of-speech tagging of transcripts from the RADIO HAÏTI-INTER radio corpus (1 300 hours) in spoken Haitian Creole. Faced with a lack of annotated data for speech, we first adapt a multilingual model (XLM-RoBERTa) through continued pre-training on the target corpus, followed by initial fine-tuning on existing Universal Dependencies treebanks. We evaluate an active learning strategy guided by a selection of training samples based on model confidence scores (random, low-confidence, high-confidence) and two fine-tuning strategies (sequential and joint). The results show that domain adaptation is crucial (+4.3 points), that the sequential approach outperforms joint fine-tuning, but that active selection does not provide a significant advantage over random sampling. We release a manually annotated sample, a language model adapted to Haitian, and a POS tagging model for transcribed spoken Haitian.

MOTS-CLÉS : apprentissage actif, annotation morphosyntaxique, étiquetage POS, créole haïtien, corpus oral, adaptation au domaine, faible ressource.

KEYWORDS: active learning, part-of-speech tagging, POS annotation, Haitian Creole, spoken corpus, domain adaptation, low-resource.

Actes des 33ème Conférence sur le Traitement Automatique des Langues Naturelles.
Volume 1 : articles scientifiques originaux, pages 121–133.

1 Introduction

Le créole haïtien (HC), parlé par plus de douze millions de locuteurices, occupe une place centrale en Haïti où il bénéficie d'un statut de co-langue officielle. Cette langue a longtemps été considérée comme une langue sévèrement sous-dotée, notamment d'un point de vue computationnel en raison de l'absence de corpus disponibles (Joshi *et al.*, 2020). Cette situation était particulièrement critique pour la modalité orale, où le manque de corpus transcrits et annotés a entravé les études linguistiques empiriques sur corpus pour les variétés spontanées. Néanmoins, le statut de la langue tend à s'améliorer grâce à quelques initiatives (Frederking *et al.*, 1997; Fattier, 1998; Valdman *et al.*, 2015; Andrus *et al.*, 2017; Valk & Alumäe, 2021). On compte également quelques ressources disponibles annotées en morphosyntaxe, à l'instar des corpus arborés Universal Dependencies de la version 2.17 (Zeman *et al.*, 2025) UD-HAITIAN-CREOLE-ADOLPHE[1] et UD-HAITIAN-CREOLE-AUTOGRAMM[2] (Kahane *et al.*, 2024), qui se limitent à des registres écrits et formels, laissant dans l'ombre les spécificités de l'oral : variations phonétiques et phonologiques, hésitations, chevauchements de parole et alternances codiques et plus largement l'analyse syntaxique des particularités de l'oral (Kahane *et al.*, 2021). Récemment, les archives radiophoniques de la station RADIO HAÏTI-INTER[3] (RHI) ont été transcrites et restructurées en corpus (Havard *et al.*, 2026). Comprenant environ 1 300 heures de parole transcrite automatiquement, le corpus marque une avancée fondamentale en offrant la première ressource orale à grande échelle pour le HC. Cependant, pour transformer cette masse de données en une ressource exploitable pour des analyses grammaticales et le développement d'outils de TAL robustes, une étape d'annotation morphosyntaxique adaptée au domaine oral reste à accomplir.

L'étiquetage grammatical de transcriptions de l'oral dans un cadre de faible ressource représente un double défi. D'une part, les modèles de langue contemporains, bien que performants sur des textes écrits standardisés, sont mal adaptés aux phénomènes de la parole et à l'écart des conventions graphiques utilisées pour représenter la réalité des productions langagières (Kumar *et al.*, 2021; Hervé *et al.*, 2022; Kanaan-Caillol *et al.*, 2025). D'autre part, l'annotation manuelle exhaustive de plusieurs millions de mots étant irréaliste, il est crucial de concevoir des méthodologies permettant de maximiser la qualité des annotations automatiques tout en optimisant l'intervention humaine. C'est précisément sur cette double problématique que cet article se penche, en proposant un pipeline complet d'étiquetage en parties du discours (*Part-Of-Speech tagging, POS tagging*) pour les transcriptions du corpus RHI. Notre approche articule plusieurs étapes pour pallier le manque de données d'entraînement : une adaptation d'un modèle de langue multilingue au domaine cible via un pré-entraînement continué sur les 13 millions de mots du corpus ; un amorçage du système d'étiquetage par affinage (*fine-tuning*) sur les seuls treebanks disponibles ; et l'implémentation d'une stratégie d'apprentissage actif guidée par les scores de confiance du modèle. Cette stratégie constitue notre hypothèse afin de sélectionner de manière incrémentale les énoncés les plus informatifs pour l'annotation manuelle, réduisant ainsi considérablement la charge de travail tout en ciblant les cas les plus ambigus pour le modèle.

Au-delà de la production d'annotations, notre travail vise également à évaluer de manière critique la fiabilité des outils automatiques dans ce contexte exigeant. Nous menons une analyse du lien entre les scores de confiance internes de l'étiqueteur et la précision réelle de ses prédictions, validant ainsi l'utilité de ces métriques non seulement pour guider l'annotation active, mais aussi pour les utilisateurices final·es du corpus qui souhaitent filtrer les segments les plus fiables pour leurs analyses. L'ensemble de la méthodologie est développée et évaluée sur un sous-ensemble de référence issu du

1. https://github.com/UniversalDependencies/UD_Haitian_Creole-Adolphe.git
2. https://github.com/UniversalDependencies/UD_Haitian_Creole-Autogramm.git
3. https://idn.duke.edu/ark:/87924/r44j0ct7h

corpus RHI, annoté manuellement et que nous mettons à disposition de la communauté.

Les contributions de ce travail sont donc multiples. Nous publions en premier lieu un échantillon du corpus Radio Haïti-Inter enrichi d'annotations POS manuelles et distribué dans un format standard (CoNLL-U)[4]. Nous fournissons ensuite un modèle de langue adapté au domaine du HC parlé, obtenu par pré-entraînement continué de XLM-RoBERTa (Conneau *et al.*, 2020, XLM)[5]. Sur cette base, nous construisons et diffusons un tagger POS spécialisé pour l'oral spontané en HC[6], ainsi qu'un second modèle pour l'annotation automatique en POS des transcriptions du créole haitien[7]. Sur le plan méthodologique, nous évaluons une stratégie d'apprentissage actif pour l'étiquetage POS en contexte de faible ressource, ainsi qu'une analyse des scores de confiance du modèle, testant leur pertinence comme levier pour la construction et l'exploitation de ressources annotées. En libérant à la fois les ressources linguistiques et les outils méthodologiques associés, ce travail vise à servir de fondation pour les recherches futures en analyse morphosyntaxique automatique des langues créoles peu dotées.

2 Corpus et ressources

Corpus Radio Haiti Inter Ce corpus est l'un des premiers corpus oraux de grande envergure pour le créole haïtien ; le premier étant l'Atlas Linguistique d'Haïti (Fattier, 1998). RADIO HAÏTI-INTER (RHI) provient des archives de la station de radio du même nom, active depuis 1935 jusqu'au début des années 2000. Cette station a joué un rôle majeur dans la valorisation de la langue du pays en étant l'un des premiers médias à diffuser principalement en créole haïtien plutôt qu'en français, couvrant une large variété de contenus (débats politiques, journaux, interviews, émissions culturelles), ainsi qu'une diversité de locuteurices (journalistes, intellectuels, citoyens ordinaires, politiciens, etc).

Le corpus brut représente environ 2 400 heures d'audio, dont 1 377 heures sont en créole haïtien. Après l'application d'un détecteur d'activité vocale et de diarisation (Plaquet & Bredin, 2023), la durée utile est de 1 165 heures. Havard *et al.* (2025) ont transcrit automatiquement les enregistrements à l'aide d'un modèle d'apprentissage autosupervisé (Baevski *et al.*, 2022) spécialisé pour le créole haïtien. Les transcriptions y sont alignées au niveau des mots et des caractères, et accompagnées de scores de confiance (de 0 à 1) calculés à partir de l'entropie de la sortie du modèle (Havard *et al.*, 2026). Ces scores permettent d'identifier les segments les plus fiables. Le corpus ainsi transcrit contient environ 13 millions de tokens.

Havard *et al.* (2026) mettent à disposition une sélection de 50 heures en maximisant la qualité des transcriptions automatiques (selon les scores de confiance) et en minimisant les chevauchements de parole. Elle couvre la même période temporelle et reflète la diversité des locuteurs et des genres du corpus complet.

Treebanks UD pour le créole haïtien Pour l'étiquetage morphosyntaxique, nous nous appuyons sur les deux seuls corpus annotés en POS, à notre connaissance, que sont les treebanks disponibles pour le créole haïtien au sein du projet Universal Dependencies (de Marneffe *et al.*, 2021, UD) :

4. https://github.com/RZiane/RZiane-TALN2026_tagger_RHI/tree/main/data
5. https://huggingface.co/Rziane/xlmr-large-kreyol-RHI
6. https://github.com/RZiane/TALN2026_tagger_RHI
7. https://huggingface.co/Rziane/xlmr-large-kreyol-RHI-pos

— UD-HAITIAN-CREOLE-ADOLPHE. Ce treebank contient 3 314 phrases (environ 300 000 tokens) tirées de sources religieuses. Les annotations proviennent d'une conversion automatique d'une annotation antérieure avec des corrections. Bien que de taille conséquente, il ne couvre qu'une modalité écrite d'un domaine spécifique (religion). De plus, le corpus est assez peu décrit : il n'y a pas de publications associées et, la seule documentation existante est dans le dépôt UD, ce qui en fait une ressource précieuse mais à considérer avec précaution.
— UD-HAITIAN-CREOLE-AUTOGRAMM. Toujours exclusivement de modalité écrite, ce treebank de 144 phrases (3 418 tokens) combine des extraits de la bible, d'un roman et de journaux. Il a été annoté manuellement selon le schéma *Surface Syntactic Universal Dependencies* (SUD) (Gerdes *et al.*, 2018) puis converti automatiquement en UD. Sa taille modeste et sa focalisation sur l'écrit formel limitent sa représentativité de l'oral spontané.

Ces deux ressources, bien qu'utiles pour amorcer l'étiquetage POS, présentent des limites importantes pour notre étude : elles sont exclusivement écrites, de registre plutôt formel, et ne reflètent pas les phénomènes spécifiques à la parole (disfluences, variations orthographiques, *code-switching*, etc). Notre corpus RHI comble ainsi un manque crucial pour l'étude du créole haïtien parlé.

Données annotées manuellement Pour évaluer et adapter notre pipeline d'étiquetage POS au domaine oral, nous avons constitué un ensemble d'annotations manuelles de référence (*gold standard*). Nous avons corrigé les transcriptions et étiqueté manuellement en POS 562 énoncés (1 728 tokens) tirés du corpus RHI. Cet échantillon couvre différents enregistrements afin de garantir une certaine diversité concernant les locuteurices, genres et périodes. Il sert à la fois à l'évaluation finale et à l'adaptation du modèle via un processus d'apprentissage actif (section 4).

3 Travaux connexes : *active learning*

L'apprentissage actif (*active learning*, AL) vise à optimiser le compromis entre coût d'annotation humaine et performance des modèles en sélectionnant les exemples jugés les plus informatifs (Settles, 2009). Pour les langues peu dotées, ce paradigme semble désormais essentiel afin de construire des jeux de données d'entraînement efficaces malgré des ressources limitées.

Les approches classiques d'AL reposent principalement sur des critères d'incertitude et de représentativité. L'échantillonnage par incertitude (*uncertainty sampling*) consiste à sélectionner les données pour lesquelles le modèle est le moins confiant, typiquement mesuré par la probabilité maximale prédite ou l'entropie de la distribution sur les étiquettes POS (Lewis, 1995). Cependant, Chaudhary *et al.* (2021) démontrent que l'incertitude seule est souvent sous-optimale pour le POS tagging. Leur analyse montre que ces méthodes échouent à cibler les confusions systématiques entre paires d'étiquettes (ex : DET/PRON) et privilégient souvent les classes majoritaires. Ils introduisent ainsi une stratégie visant explicitement à réduire la confusion en identifiant les types lexicaux ambigus. Par ailleurs, la mauvaise calibration des modèles de langue (Transformers), qui ont tendance à être sur-confiants, peut limiter la fiabilité de scores d'incertitude (Desai & Durrett, 2020).

En complément des critères d'incertitude, des mesures de représentativité cherchent à éviter une sélection excessive d'exemples atypiques ou redondants, par exemple via des mesures de diversité ou de couverture de l'espace des données non annotées (Huang *et al.*, 2014; Estève *et al.*, 2025).

Une étude récente menée par Liu *et al.* (2025) sur 12 familles de langues compare l'efficacité de

l'AL face à l'échantillonnage aléatoire et à l'apprentissage en contexte (*In-Context Learning*) via des modèles propriétaires. Leurs résultats indiquent que pour les communautés ne souhaitant pas partager leurs données avec des API tierces, l'AL permet d'atteindre des performances raisonnables avec un volume réduit. Cette étude confirme que l'AL surpasse l'échantillonnage aléatoire en termes de rapidité d'apprentissage.

L'AL semble ainsi permettre d'atteindre un plateau de performance plus rapidement, validant son usage comme levier d'accélération pour l'annotation de langues peu dotées comme le créole haïtien.

4 Méthodologie

Cette partie présente le protocole expérimental en détail, à partir de la spécialisation d'un modèle pré-entraîné vers l'utilisation d'un score de confiance afin d'échantillonner les lots d'entraînements, en passant par l'adaptation du modèle pré-entraîné pour la tâche de classification en POS.

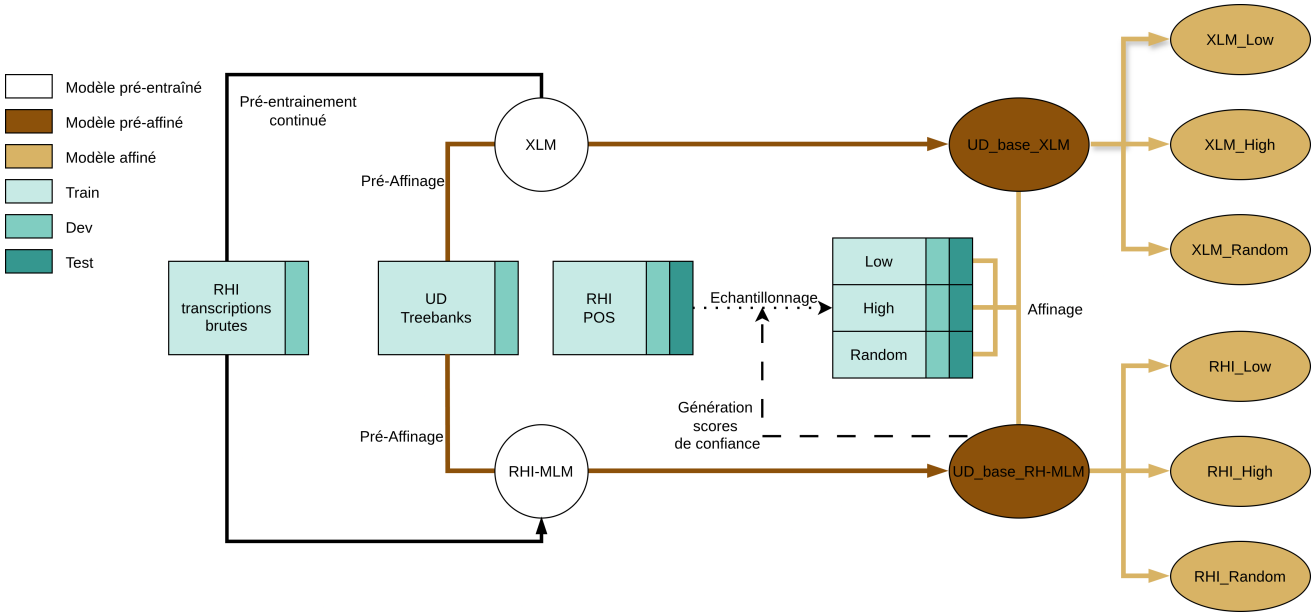

FIGURE 1 – Schéma récapitulatif du pipeline expérimental.

4.1 Modèles et score de confiance

Nous décrivons ici les étapes de préparation du modèle, incluant son adaptation au domaine cible, sa spécialisation à la tâche de POS tagging, et la génération des scores de confiance qui seront utilisés pour l'apprentissage actif.

Adaptation par MLM du modèle XLM au corpus cible Le point de départ de notre pipeline expérimental est une phase d'adaptation du modèle de langue multilingue XLM au corpus cible RADIO HAÏTI-INTER (Ramponi & Plank, 2020; Grobol *et al.*, 2022). Cette étape de *pré-entraînement continué* (*continued pretraining*) vise à ajuster les représentations du modèle — initialement entraîné sur 100 langues (à l'exclusion du créole haïtien) — aux spécificités lexicales, morphologiques,

phonologiques et graphiques du créole haïtien telles qu'attestées dans les transcriptions du corpus radiophonique. Concrètement, nous utilisons l'objectif de modèle de langue masqué (MLM) où 15% des tokens sont masqués aléatoirement et le modèle doit les reconstruire. L'entraînement est effectué pendant 3 époques sur l'ensemble du corpus RADIO HAÏTI-INTER (949 712 mots d'entraînement, 237 429 de validation), avec un taux d'apprentissage de 5×10^{-5}, une taille de batch 16 et l'activation du *gradient checkpointing* pour optimiser l'utilisation mémoire. Cette adaptation permet au modèle d'intégrer les régularités distributionnelles du RADIO HAÏTI-INTER, réduisant ainsi l'écart entre le pré-entraînement multilingue initial et le domaine cible.

Pré-affinage sur les treebanks UD du créole haïtien Le modèle adapté par MLM est ensuite spécialisé pour la tâche d'étiquetage morphosyntaxique via un pré-affinage avec les deux treebanks disponibles en créole haïtien dans le projet Universal Dependencies : UD-HAITIAN-CREOLE-AUTOGRAMM et UD-HAITIAN-CREOLE-ADOLPHE. Ces ensembles ont été sélectionnés pour leur proximité linguistique avec notre corpus d'étude. L'étape ici décrite spécialise le modèle de langue générique en un classifieur de tokens capable d'attribuer des étiquettes POS pour le créole haïtien, sans adaptation à l'analyse en parties du discours des transcriptions du corpus cible pour le moment. L'entraînement[8] du modèle de classification, initialisé avec les poids du modèle MLM, utilise une perte d'entropie croisée, l'optimiseur AdamW avec décroissance de poids (0.01), un taux d'apprentissage de 2×10^{-5}, et est exécuté sur 20 époques selon la configuration d'un *early stopping* basé sur la patience de 10 époques.

Génération des scores de confiance et application au corpus cible Le modèle pré-adapté est ensuite appliqué au corpus RADIO HAÏTI-INTER pour produire des étiquettes POS prédites accompagnées de *scores de confiance*. Pour chaque token, le vecteur de logits en sortie du modèle est converti en distribution de probabilités obtenue via une fonction softmax, produisant un vecteur $\mathbf{p}$ où chaque élément p_k représente la probabilité attribuée à la classe POS k. Le score de confiance est calculé comme l'entropie normalisée de cette distribution : $c = 1 - \frac{H(\mathbf{p})}{\log(N)}$, où $H(\mathbf{p}) = -\sum_{k=1}^{N} p_k \log(p_k)$ est l'entropie de Shannon de la distribution et N le nombre total de classes POS. Cette métrique, comprise entre 0 (incertitude maximale) et 1 (certitude absolue), reflète le degré de confiance du modèle dans sa prédiction pour chaque token. Plus précisément, une valeur proche de 1 indique que la distribution de probabilités est fortement concentrée sur une seule classe, tandis qu'une valeur proche de 0 signale une répartition quasi-uniforme entre les différentes classes, traduisant une incertitude élevée.

Le modèle pré-finetuné est appliqué à l'ensemble de notre corpus annoté manuellement, qui comprend 562 énoncés (1728 tokens) au format CoNLL-U. Pour chaque token t_{ij} (token j de l'énoncé i), le modèle produit :

— Une étiquette POS prédite $\hat{y}_{ij}$;

— Un score de confiance $c_{ij} = 1 - \frac{H(\text{softmax}(\mathbf{l}_{ij}))}{\log(C)}$, où C est le nombre de classes POS.

Pour valider la pertinence des scores de confiance comme indicateur de fiabilité, nous avons mesuré leur corrélation avec la justesse des prédictions sur l'ensemble de référence annoté manuellement. La corrélation de Pearson est modérée mais significative ($r = 0,395$, $p < 0,001$). L'examen par déciles montre que la précision passe de 54,3 % dans le premier décile (confiance moyenne = 0,76) à

8. L'architecture employée est `AutoModelForTokenClassification` de la bibliothèque `transformers` d'HuggingFace (Wolf *et al.*, 2020).

96,1 % dans le septième (confiance = 1,00). Cette relation varie selon les catégories grammaticales (par exemple, $r = 0,99$ pour NUM, $r = 0,20$ non significatif pour CCONJ), mais valide globalement l'usage des scores comme critère de sélection pour la construction des lots expérimentaux.

4.2 Stratégie de sélection basée sur la confiance pour l'apprentissage actif

Cette sous-section présente la méthodologie de sélection active des échantillons à annoter, basée sur les scores de confiance, ainsi que la construction des lots expérimentaux pour évaluer cette stratégie.

Sélection des échantillons La stratégie de sélection pour l'apprentissage actif repose sur une métrique de confiance relative qui s'appuie sur la distribution globale des scores de confiance dans le corpus. Cette approche permet d'identifier les énoncés pour lesquelles le modèle présente le plus d'incertitude, ciblant ainsi les échantillons potentiellement les plus informatifs pour l'affinage du modèle et à privilégier pour l'annotation humaine. Le processus se décompose en deux étapes principales. Premièrement, nous calculons un seuil de référence τ défini comme la médiane des scores de confiance de tous les tokens du corpus $\mathcal{T} = \{t_{ij} : \forall i, j\}$, soit $\tau = \text{Médiane}(\{c_{ij} : t_{ij} \in \mathcal{T}\})$ où j sont les tokens de l'énoncé i.

Le choix de la médiane plutôt que de la moyenne confère à notre métrique une robustesse statistique face aux valeurs extrêmes observées dans les données, garantissant que le seuil n'est pas indûment influencé par quelques tokens présentant des scores exceptionnellement faibles ou élevés. Ce seuil représente ainsi le niveau de confiance médian du modèle sur l'ensemble du corpus.

Deuxièmement, pour chaque phrase s_i composée de n_i tokens, nous calculons un ratio d'incertitude $\mathcal{C}(s_i) = \frac{|\{t_{ij} \in s_i : c_{ij} < \tau\}|}{n_i}$, qui représente la proportion de tokens dans la phrase dont la confiance est inférieure au seuil médian global. Une valeur élevée de $\mathcal{C}(s_i)$ indique que le modèle est incertain sur une grande partie des tokens de la phrase, signalant un échantillon potentiellement difficile à annoter automatiquement mais informatif pour l'apprentissage.

La métrique $\mathcal{C}(s_i)$ présente plusieurs propriétés avantageuses. Sa normalisation par la longueur de la phrase la rend indépendante du nombre de tokens, permettant ainsi des comparaisons plus équitables entre énoncés de tailles différentes. Son interprétation est intuitive : elle varie continûment entre 0 et 1, où 0 indique que tous les tokens sont « confiants » (au-dessus du seuil médian) et 1 qu'aucun token n'est « confiant » (tous en dessous du seuil). Cette interprétation facilite la sélection manuelle ou automatique des échantillons. La sensibilité contextuelle est assurée par l'adaptation automatique du seuil τ aux caractéristiques du corpus, contrairement à un seuil fixe qui pourrait ne pas être approprié pour différents domaines (notre corpus étant composé de diverses sources d'enregistrements allant du podcast en studio à l'interview de rue). Enfin, l'utilisation conjointe de la médiane et d'une proportion confère à la métrique une robustesse face aux valeurs aberrantes et aux variations mineures des scores de confiance des tokens pris individuellement.

Application à la stratification du corpus La métrique de confiance relative $\mathcal{C}(s_i)$ permet d'ordonner les énoncés selon leur degré global d'incertitude. En triant les énoncés par ordre décroissant de cette métrique, les échantillons les plus incertains – c'est-à-dire ceux pour lesquels le modèle présente la plus forte proportion de tokens peu confiants – sont placés en tête de liste. Cette séquence ordonnée sert de base à la construction de trois sous-ensembles distincts : un sous-ensemble à faible

confiance regroupant les énoncés aux valeurs $\mathcal{C}(s_i)$ les plus élevées (échantillons les plus incertains), un sous-ensemble à haute confiance constitué des énoncés aux valeurs les plus basses (échantillons les plus certains), et un sous-ensemble aléatoire obtenu par tirage aléatoire indépendant de la métrique, qui sert de référence. Cette stratification permet de tester l'hypothèse selon laquelle l'entraînement sur des données incertaines peut conduire à de meilleures performances, car ces données contiennent potentiellement des cas difficiles qui, une fois annotés, apportent une information plus riche au modèle. Toutefois, une telle sélection basée uniquement sur la proportion de tokens peu confiants pourrait introduire un biais lié à la longueur des énoncés : une phrase courte avec un seul token peu confiant obtiendrait une valeur $\mathcal{C}(s_i)$ de 1,0, tandis qu'une phrase longue avec plusieurs tokens peu confiants mais aussi de nombreux tokens confiants pourrait voir sa valeur diluée. Bien que la métrique soit normalisée par la longueur, une relation systématique entre incertitude et longueur demeure plausible (les énoncés longs, souvent plus complexes, pourraient présenter davantage de tokens incertains). Pour isoler l'effet de la confiance et éviter que les différences observées ne soient attribuables à des disparités de longueur, nous procédons à une stratification préalable par longueur de phrase.

Stratification par longueur des énoncés Afin d'éviter qu'un éventuel biais lié à la longueur des énoncés n'affecte la sélection, nous stratifions préalablement le corpus en cinq groupes de longueur ($\mathcal{G}_1$ à $\mathcal{G}_5$ selon le nombre de tokens, avec les bornes [1–10], [11–20], [21–30], [31–40] et 41+), choisis pour équilibrer les effectifs. La sélection par confiance est ensuite appliquée indépendamment dans chaque groupe.

Construction des lots expérimentaux Le corpus de 562 énoncés est d'abord partitionné en ensembles d'entraînement, de développement et de test (60% – 20% – 20%) en préservant la distribution des groupes de longueur. Pour chaque groupe de longueur $\mathcal{G}_k$ au sein d'un même ensemble, nous répartissons aléatoirement un tiers des énoncés dans le lot "aléatoire" (`rand`). Parmi les deux tiers restants, la moitié des énoncés les plus incertains ($\mathcal{C}(s_i)$ le plus élevé) constituent le lot "faible confiance" (`low`), et les autres le lot "haute confiance" (`high`). Cette opération est répétée pour les ensembles d'entraînement, de développement et de test. On obtient ainsi neuf lots expérimentaux (trois par ensemble), chacun contenant un mélange équilibré de phrases de toutes longueurs. Les lots de développement sont ensuite concaténés pour servir de validation unique. Les lots d'entraînement comptent environ 100 énoncés chacun, ceux de développement et de test environ 37.

4.3 Protocole expérimental d'affinage

Six modèles identiques (initialisés avec les mêmes poids) sont d'abord affinés séparément avec $\text{Train}_{\text{rand}}$, $\text{Train}_{\text{low}}$, puis $\text{Train}_{\text{high}}$ en utilisant les plongements lexicaux des modèles XLM et RHI-MLM. Cette approche permet d'isoler l'effet de la composition des données d'entraînement sur les performances finales. Les hyperparamètres sont fixés identiquement pour les trois conditions afin d'assurer la comparabilité des résultats : un taux d'apprentissage de 5×10^{-5}, une taille de batch de 16, un maximum de 50 époques avec un arrêt anticipé basé sur une patience de 5 époques sur la perte de validation, l'optimiseur AdamW avec une décroissance de poids de 0,01, et un planificateur de taux d'apprentissage linéaire avec échauffement de 10%.

Enfin chaque modèle est affiné cinq fois à partir de *seeds* d'initialisation aléatoire différentes dans le but d'assurer une robustesse des observations.

Dans la lignée des travaux de Phang *et al.* (2019) et Weller *et al.* (2022), nous comparons deux stratégies d'affinage pour l'apprentissage par transfert. La première, dite affinage séquentiel, correspond à la méthodologie principale décrite en section 4.1. La seconde, l'affinage joint, consiste à combiner les deux domaines (texte écrit et transcriptions de l'oral) en concaténant les jeux de données UD et les annotations gold du corpus RHI au sein d'un même ensemble d'entraînement.

Les douzes modèles résultants de cette série d'entraînement sont évalués sur chaque échantillon de test.

5 Résultats

Notre évaluation permet de mesurer l'efficacité de l'apprentissage actif guidé par les scores de confiance pour l'annotation en POS du créole haïtien. L'analyse révèle des effets nuancés de cette approche, tout en mettant en lumière l'importance d'autres facteurs méthodologiques secondaires parmi nos expérimentations.

L'hypothèse selon laquelle la sélection active d'énoncés basée sur les scores de confiance permettrait d'améliorer significativement les performances avec un volume réduit d'annotations manuelles ne trouve qu'un soutien partiel dans nos résultats. En effet, les trois stratégies de sélection : aléatoire (`random`), basée sur les énoncés à faible confiance (`low`), et basée sur les énoncés à haute confiance (`high`), produisent des performances globales quasiment équivalentes sur l'ensemble complet de test.

Pour le modèle adapté au domaine via pré-entraînement continué (RHI-MLM) et utilisant un affinage séquentiel, les précisions moyennes sont respectivement de $92,4$ ($\pm 0,5$), $92,3$ ($\pm 0,3$) et $92,1$ ($\pm 0,3$) (tableau 1). Ces différences, de l'ordre de quelques millièmes, ne permettant pas de conclure à un avantage significatif d'une stratégie particulière.

Contrairement aux effets modestes de la stratégie de sélection, l'adaptation préalable du modèle au domaine cible via pré-entraînement continu produit une amélioration substantielle et systématique des performances. Le modèle XLM-RoBERTa-large adapté au corpus RHI (RHI-MLM) atteint une précision de base de $86 \pm 0,4$ après spécialisation initiale sur les treebanks UD, contre seulement $81,7 \pm 1$ pour le modèle non adapté (XLM, tableau 1). Cet écart de $+4,3$ points démontre l'importance de réduire la divergence entre le pré-entraînement multilingue et les caractéristiques spécifiques du domaine cible.

Modèle	Test Random	Test Low	Test High	Test All
RHI-MLM				
UD_base_RHI-MLM	85,2 ± 0,6	86,8 ± 0,9	86,1 ± 0,3	86,0 ± 0,4
RHI-MLM_random	92,1 ± 0,6	**92,9 ± 0,5**	92,1 ± 0,6	**92,4 ± 0,5**
RHI-MLM_low	91,6 ± 0,2	**92,9 ± 0,6**	**92,4 ± 0,4**	92,3 ± 0,3
RHI-MLM_high	**92,7 ± 0,6**	92,8 ± 0,2	91,2 ± 0,7	92,1 ± 0,3
XLM				
UD_base_XLM	81,6 ± 1,4	80,2 ± 1,0	82,7 ± 0,7	81,7 ± 1,0
XLM_random	89,2 ± 0,7	**90,5 ± 0,8**	**89,2 ± 0,6**	**89,6 ± 0,5**
XLM_low	87,7 ± 0,8	88,9 ± 0,7	87,8 ± 0,7	88,2 ± 0,5
XLM_high	**89,8 ± 0,3**	90,4 ± 0,7	88,5 ± 0,3	89,4 ± 0,4

TABLE 1 – Moyenne et écart-type (sur 5 seeds) des précisions pour chaque modèle et chaque jeu de test en pourcentage – affinage séquentiel.

Modèle	Test Random	Test Low	Test High	Test All
RHI-MLM				
UD_base_RHI-MLM	85,2 ± 0,6	86,8 ± 0,9	86,1 ± 0,3	86,0 ± 0,4
RHI-MLM_random	89,9 ± 1,2	91,7 ± 0,6	**90,5 ± 0,8**	90,7 ± 0,7
RHI-MLM_low	90,3 ± 0,4	**91,9 ± 0,8**	90,4 ± 0,4	90,8 ± 0,2
RHI-MLM_high	**91,1 ± 0,7**	91,8 ± 0,6	**90,5 ± 0,6**	**91,1 ± 0,5**
XLM				
UD_base_XLM	81,6 ± 1,4	80,2 ± 1,0	82,7 ± 0,7	81,7 ± 1,0
XLM_random	88,4 ± 1,0	**89,4 ± 0,7**	**88,1 ± 0,4**	88,5 ± 0,5
XLM_low	88,4 ± 1,1	88,6 ± 0,4	87,5 ± 0,5	88,1 ± 0,6
XLM_high	**89,9 ± 0,7**	88,7 ± 0,6	87,9 ± 0,4	**88,7 ± 0,3**

TABLE 2 – Moyenne et écart-type (sur 5 seeds) des précisions pour chaque modèle et chaque jeu de test en pourcentage – affinage joint.

Cette adaptation préalable influence également l'efficacité relative des stratégies de sélection. Pour le modèle non adapté (XLM), la stratégie `low` produit des performances globales légèrement inférieures ($88,2 \pm 0,5$) à celles des stratégies `random` ($89,6 \pm 0,5$) et `high` ($89,4 \pm 0,4$). Cette différence suggère que l'efficacité de l'apprentissage actif basé sur les scores de confiance pourrait dépendre de la qualité des représentations sous-jacentes du modèle, elle-même améliorée par l'adaptation au domaine.

La comparaison entre les stratégies d'affinage séquentiel et joint révèle un avantage systématique de l'approche séquentielle. Pour le modèle adapté (RHI-MLM), l'approche séquentielle permet d'atteindre une précision moyenne de $92,4$ ($\pm 0,5$) avec sélection aléatoire, contre $90,7$ ($\pm 0,7$) avec l'affinage joint (tableaux 1 et 2). Cet écart de $+1,7$ point confirme l'intérêt d'une spécialisation progressive : un premier affinage sur des données écrites proches dans la langue cible (treebanks UD) pour acquérir les compétences de base en étiquetage morphosyntaxique, suivi d'un second affinage sur un sous-ensemble ciblé du domaine oral pour l'adapter aux spécificités de ce type de données.

Pour le modèle non adapté (XLM), la tendance est similaire mais moins marquée, avec des performances de $89,6$ ($\pm 0,5$) en séquentiel contre $88,5$ ($\pm 0,5$) en affinage joint pour la sélection aléatoire. Cette cohérence dans les résultats suggère que la séparation des phases d'apprentissage permet au modèle de mieux intégrer les connaissances successives, peut-être avec moins d'interférences entre les distributions de données différentes.

6 Conclusion

Nos résultats mettent en évidence une hiérarchie dans l'impact des différentes composantes méthodologiques. L'adaptation préalable au domaine par pré-entraînement continu constitue le facteur le plus déterminant dans les performances finales, apportant une amélioration substantielle de plus de 4 points de pourcentage. Le choix d'une stratégie d'affinage séquentiel plutôt que cumulative offre un avantage complémentaire, bien que plus modeste, de l'ordre de 1 à 2 points. En revanche, la stratégie de sélection active basée sur les scores de confiance ne démontre pas d'avantage par rapport à un échantillonnage aléatoire dans notre configuration expérimentale. Néanmoins, la corrélation positive entre les scores de confiance du modèle et la précision des annotations permet d'offrir une meilleure vision quant aux prédictions du modèle et d'isoler les annotations automatique les plus fiables afin de trier parmi la masse de données disponibles.

Ces observations invitent à reconsidérer la priorisation des efforts dans des contextes similaires de faible ressource : l'investissement dans l'adaptation préalable du modèle au domaine cible et dans une stratégie d'apprentissage progressive apparaît plus rentable que la mise en œuvre de mécanismes complexes de sélection active basés uniquement sur les scores de confiance telle que nous l'avons testée. Par ailleurs, conscient de la quantité modeste de données sur laquelle les expériences ont été menées, un changement d'échelle pourrait être associé à des conclusions différentes.

Remerciements

Ce travail a bénéficié du soutien partiel de l'Agence Nationale de la Recherche, via le projet SynPaX (ANR-23-CE23-0017-01). Nous remercions également la Direction du Système d'Information de

l'Université de Caen pour l'accès aux ressources de calcul ayant permis l'entraînement des modèles.

Références

ANDRUS T., BILLS A., CONNERS T., CRABB E. S., DUBINSKI E., FISCUS J. G., GILLIES B., HARPER M., HAZEN T., HEFRIGHT B. *et al.* (2017). Iarpa babel haitian creole language pack iarpa-babel201b-v0. 2b.

BAEVSKI A., HSU W.-N., XU Q., BABU A., GU J. & AULI M. (2022). data2vec : A General Framework for Self-supervised Learning in Speech, Vision and Language. In *Proceedings of the 39th International Conference on Machine Learning*, p. 1298–1312 : PMLR.

CHAUDHARY A., ANASTASOPOULOS A., SHEIKH Z. & NEUBIG G. (2021). Reducing Confusion in Active Learning for Part-Of-Speech Tagging. *Transactions of the Association for Computational Linguistics*, **9**. DOI : 10.1162/tacl_a_00350.

CONNEAU A., KHANDELWAL K., GOYAL N., CHAUDHARY V., WENZEK G., GUZMÁN F., GRAVE E., OTT M., ZETTLEMOYER L. & STOYANOV V. (2020). Unsupervised Cross-lingual Representation Learning at Scale. arXiv :1911.02116 [cs], DOI : 10.48550/arXiv.1911.02116.

DE MARNEFFE M.-C., MANNING C. D., NIVRE J. & ZEMAN D. (2021). Universal Dependencies. *Computational Linguistics*, **47**(2), 255–308. DOI : 10.1162/coli_a_00402.

DESAI S. & DURRETT G. (2020). Calibration of Pre-trained Transformers. arXiv :2003.07892 [cs], DOI : 10.48550/arXiv.2003.07892.

ESTÈVE L., MARNEFFE M.-C. D., MELNIK N., SAVARY A. & KANISHCHEVA O. (2025). A survey of diversity quantification in natural language processing : The why, what, where and how. arXiv :2507.20858 [cs], DOI : 10.48550/arXiv.2507.20858.

FATTIER D. (1998). *Contribution à l'étude de la genèse d'un créole : l'atlas linguistique d'Haïti, cartes et commentaires*. Thèse de doctorat, Aix-Marseille 1.

FREDERKING R., RUDNICKY A. & HOGAN C. (1997). Interactive speech translation in thediplomatproject. In *Spoken Language Translation w orkshop at the 3 5th Meeting of the Association forComputational Linguistics, AC L-97. Madrid, Spain*.

GERDES K., GUILLAUME B., KAHANE S. & PERRIER G. (2018). Sud or surface-syntactic universal dependencies : An annotation scheme near-isomorphic to ud. In T. LYNN & S. SCHUSTER, Éds., *Universal Dependencies Workshop 2018*, Brussels, Belgium. DOI : 10.18653/v1/W18-6008.

GROBOL L., REGNAULT M., ORTIZ SUAREZ P., SAGOT B., ROMARY L. & CRABBÉ B. (2022). BERTrade : Using contextual embeddings to parse Old French. In N. CALZOLARI, F. BÉCHET, P. BLACHE, K. CHOUKRI, C. CIERI, T. DECLERCK, S. GOGGI, H. ISAHARA, B. MAEGAARD, J. MARIANI, H. MAZO, J. ODIJK & S. PIPERIDIS, Éds., *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, p. 1104–1113, Marseille, France : European Language Resources Association.

HAVARD W., ZIANE R., MENCLÉ M., COAVOUX M., LECOUTEUX B. & EMMANUEL S. (2026). Radio Haiti-Inter : a large-scale annotated corpus of spoken Haitian Creole. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, Palma, Spain. Accepté.

HAVARD W. N., GOVAIN R., LECOUTEUX B. & SCHANG E. (2025). Self-Supervised Models of Speech Processing for Haitian Creole. In *Interspeech 2025*, p. 4018–4022. DOI : 10.21437/Interspeech.2025-1852.

HERVÉ N., PELLOIN V., FAVRE B., DARY F., LAURENT A., MEIGNIER S. & BESACIER L. (2022). Using ASR-generated text for spoken language modeling. In A. FAN, S. ILIC, T. WOLF & M. GALLÉ, Éds., *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, p. 17–25, virtual+Dublin : Association for Computational Linguistics. DOI : 10.18653/v1/2022.bigscience-1.2.

HUANG S.-J., JIN R. & ZHOU Z.-H. (2014). Active Learning by Querying Informative and Representative Examples. *IEEE transactions on pattern analysis and machine intelligence*, **36**(10), 1936–1949. DOI : 10.1109/TPAMI.2014.2307881.

JOSHI P., SANTY S., BUDHIRAJA A., BALI K. & CHOUDHURY M. (2020). The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. DOI : 10.18653/v1/2020.acl-main.560.

KAHANE S., CARON B., STRICKLAND E. & GERDES K. (2021). Annotation guidelines of UD and SUD treebanks for spoken corpora. p. pp. 35. Association for Computational Linguistics.

KAHANE S., PIERRE-LOUIS C., JAGODZIŃSKA S. & SAVARY A. (2024). The first Haitian Creole treebank. Published : Peer reviewed poster in the 2nd UniDive Workshop.

KANAAN-CAILLOL L., DUGUA C. & GERSTENBERG A. (2025). *Représenter la parole : une dimension fondamentale de la linguistique*, In L. KANAAN-CAILLOL, C. DUGUA & A. GERSTENBERG, Éds., *Représenter la parole : Apports à une dimension fondamentale de la linguistique*, p. 1–8. De Gruyter.

KUMAR A., NARAYANAN SUNDARARAMAN M. & VEPA J. (2021). What BERT based language model learns in spoken transcripts : An empirical study. In J. BASTINGS, Y. BELINKOV, E. DUPOUX, M. GIULIANELLI, D. HUPKES, Y. PINTER & H. SAJJAD, Éds., *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, p. 322–336, Punta Cana, Dominican Republic : Association for Computational Linguistics. DOI : 10.18653/v1/2021.blackboxnlp-1.25.

LEWIS D. D. (1995). A sequential algorithm for training text classifiers : corrigendum and additional data. *ACM SIGIR Forum*, **29**(2), 13–19. DOI : 10.1145/219587.219592.

LIU Z., JASBI M., GRANT C., SAGAE K. & PRUD'HOMMEAUX E. (2025). What data should I include in my POS tagging training set ? p. 8439–8455, Suzhou, China : Association for Computational Linguistics. DOI : 10.18653/v1/2025.findings-emnlp.448.

PHANG J., FÉVRY T. & BOWMAN S. R. (2019). Sentence encoders on stilts : Supplementary training on intermediate labeled-data tasks.

PLAQUET A. & BREDIN H. (2023). Powerset multi-class cross entropy loss for neural speaker diarization. In *Interspeech 2023*, p. 3222–3226. DOI : 10.21437/Interspeech.2023-205.

RAMPONI A. & PLANK B. (2020). Neural unsupervised domain adaptation in NLP—A survey. In D. SCOTT, N. BEL & C. ZONG, Éds., *Proceedings of the 28th International Conference on Computational Linguistics*, p. 6838–6855, Barcelona, Spain (Online) : International Committee on Computational Linguistics. DOI : 10.18653/v1/2020.coling-main.603.

SETTLES B. (2009). *Active Learning Literature Survey*. Computer Sciences Technical Report 1648, University of Wisconsin–Madison.

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.

VALK J. & ALUMÄE T. (2021). Voxlingua107 : a dataset for spoken language recognition.

WELLER O., SEPPI K. & GARDNER M. (2022). When to use multi-task learning vs intermediate fine-tuning for pre-trained encoder transfer learning. In S. MURESAN, P. NAKOV & A. VILLA-

VICENCIO, Éds., *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2 : Short Papers)*, p. 272–282, Dublin, Ireland : Association for Computational Linguistics. DOI : 10.18653/v1/2022.acl-short.30.

WOLF T., DEBUT L., SANH V., CHAUMOND J., DELANGUE C., MOI A., CISTAC P., RAULT T., LOUF R., FUNTOWICZ M., DAVISON J., SHLEIFER S., VON PLATEN P., MA C., JERNITE Y., PLU J., XU C., LE SCAO T., GUGGER S., DRAME M., LHOEST Q. & RUSH A. (2020). Transformers : State-of-the-art natural language processing. In Q. LIU & D. SCHLANGEN, Éds., *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing : System Demonstrations*, p. 38–45, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.emnlp-demos.6.

ZEMAN D. *et al.* (2025). Universal dependencies 2.17. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL).