Radio Haiti-Inter: a Large-Scale Annotated Corpus of Spoken Haitian Creole

William N. Havard[1,2], **Rayan Ziane**[1], **Mélissa Menclé**[1], **Maximin Coavoux**[2], **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
<first>.<last>@{univ-orleans.fr, univ-grenoble-alpes.fr}

Abstract

We present the first large-scale corpus of spoken Haitian Creole (*Kreyòl*), namely RADIO HAITI-INTER. The corpus was constructed using automatic speech recognition (ASR) with a state-of-the-art model specifically dedicated to *Kreyòl*. In addition to transcriptions, we provide part-of-speech (POS) tags, as well as time-aligned transcripts and confidence scores, enabling users to select the most reliable segments for their research. We conduct a manual evaluation of both the transcription quality and POS tagging accuracy to assess the reliability of the resource we present. To enable high-quality research with the resource we introduce, we are releasing 50 hours, comprising both the audios and attached annotations, drawn from the highest-quality segments. This corpus represents an invaluable resource for advancing the study of *Kreyòl*, with potential applications in phonetics, phonology, morphology, syntax, as well as the study of code-switching and code-mixing. As the recordings cover a large span of years, the corpus we introduce is also suited to micro-diachronic studies of *Kreyòl*.

Keywords: Haitian Creole, speech corpus, corpus linguistics

1. Introduction

Haitian Creole (ISO 639-3: hat), called *Kreyòl* by its speakers, is a French-lexifier (see below) creole spoken by around 12 million people in Haiti and diaspora communities in the Caribbean and North America. It has co-official status with French in Haiti and a standardised orthography used in education and the media (the official alphabet is based on Latin and is mainly phonemic). The label *French-lexifier* comes from the fact that the major part of *Kreyòl*'s words find their etymon in French. But regarding its grammar, it differs significatively from French: number (-yo) and definiteness (-la, -a, -nan, -an) are post-nominal elements; verbs do not inflect and pre-verbal markers (te, ap, pral, etc.) encode tense, mood and aspect. *Kreyòl* is spoken in a context of diglossia and still faces challenges in achieving parity with French, which is considered a more prestigious variety, notwithstanding efforts within Haiti to grant it an equal status.

Despite being one of the world's major languages – Eberhard et al. (2024) has ranked *Kreyòl* among the top 100 most widely spoken languages since 1999 at least – *Kreyòl* remains acutely understudied, both linguistically and computationally. Linguistic research has been particularly hindered by the scarcity of accessible recordings, and when they existed, by the unavailability of transcriptions, as analyses mainly depend on the availability of transcribed data. From a computational perspective, Joshi et al. (2020) classified *Kreyòl* as a CLASS-0 language[1] — along with 2,000+ other languages — indicating that it is "*still ignored in the aspect of language technologies*" with "*exceptionally limited resources*" and "*virtually no unlabeled data to use*".

This characterisation has recently been challenged by several works, both in the textual and spoken domains. First, regarding the textual domain, several initiatives, such as Lent et al. (2022) or Lent et al. (2024) have aimed to survey and build language resources for the major creole languages of the world, and create benchmarks allowing to test computational models. Regarding the spoken domain, up until very recently, there were no state-of-the-art speech recognition models comparable to those available for English or French, that would enable automatic transcription at scale for *Kreyòl*. However, Havard et al. (2025) recently introduced WAV2VEC2 and DATA2VEC-based models pre-trained purely on *Kreyòl* data, and further fine-tuned on a speech recognition task.

However, despite these advances, the resulting progress has not yet been fully integrated into linguistic research, leaving the field of linguistic analysis for *Kreyòl* lagging behind technological developments. The corpus we introduce in this paper aims to fill this gap. In this work, we introduce the RADIO HAÏTI-INTER corpus, which comprises more than 1.3k hours of automatically transcribed speech. The recordings originate from the RADIO HAÏTI-INTER radio station (see section 5), which played a pivotal role in Haiti both from a sociolinguistic perspective, as one of the first major media outlets to broadcast primarily in *Kreyòl* rather than

[1] See https://microsoft.github.io/linguisticdiversity/assets/lang2tax.txt (last accessed: 2025-10-10).

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 3083–3093
11–16 May 2026.

French,[2] and from a social perspective, by advocating for democracy. Its archives document not only key historical moments but also everyday speech, debates, and cultural life across several decades, from the 1970s to the early 2000s. The recordings capture a wide diversity of speakers – from journalists & intellectuals, to ordinary citizens – spanning multiple social backgrounds and linguistic registers.

Our work sheds new light on this invaluable archive material by making them accessible through automatic transcription. The RADIO HAÏTI-INTER corpus constitutes the first large-scale reference oral resource for *Kreyòl*, enabling linguistic investigations across multiple levels of analysis, including phonetics, syntax, and semantics, as well as micro-diachronic studies, code-switching/mixing phenomena, disfluency patterns, and more.

Contributions. To summarise, our contributions go as follows: we introduce the largest ever corpus of spoken *Kreyòl*, consisting of 1.3k hours of automatically transcribed speech, which we also tag for part-of-speech (POS). To ensure high-quality research with the resource we produce, we sampled 50h that are made freely available to the general public (Table 1).[3] It is accessible on Zenodo (zenodo.org), under the following DOI: `10.5281/zenodo.17818122`.

In order to be usable by linguists and creolists alike, we also provide word level and character level alignments, as well as confidence scores, allowing to sample the transcriptions the speech recognition model was most confident about. The data is distributed in open standard formats (CSVs, JSONs) and conversion scripts are provided to convert them to ELAN's EAF and Praat's TextGrids formats.

2. Related Work

Spoken Creole Corpora. Fully transcribed spoken corpora for Creole languages remain scarce and, when available, are generally modest in size. For English-lexifier Creoles, the only resource we are aware of is the NAIJASYNCOR corpus, containing about 20h of speech (Caron, 2020). For French-lexifier Creoles, resources are somewhat richer: a fully transcribed corpus of Mauritian Creole has recently been released (Tonjes Veenstra, 2024). In the Caribbean, *Kreyòl* is comparatively well served: a US-funded effort produced a larger corpus of approximately 80h (after silence removal, Andrus, Tony et al., 2017) consisting mainly of speech recorded over the phone and some read speech, complementing an earlier collection of roughly 20h

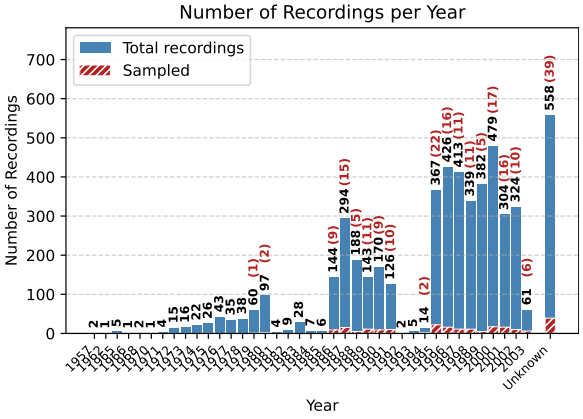

Figure 1: Distribution of number of recordings per year for the whole corpus (**blue**) and for the sampled recordings (**red**), the latter corresponding to the highest quality material, that will be openly distributed, consisting of 50h of speech.

(Carnegie Mellon University, 2010) that only consisted of read speech. For other regional varieties, only a small Guadeloupean/Martinican Creole corpus (Zribi-Hertz et al., 2012) exists. Beyond these initiatives, we are not aware of any other fully transcribed Creole speech corpus.

Annotated Creole Corpora. Creole languages remain severely under-resourced in terms of annotated corpora, and POS tagged datasets have been especially rare (Lent et al., 2022). Only a few projects have developed POS-annotated corpora for Creoles, often on a limited scale. One early example is the Gulf of Guinea Creole Corpora, which compiled written texts and transcribed spoken data for four Portuguese-lexified Creoles (Santome, Angolar, Principense, Fa d'Ambô) and annotated them with a custom POS tagset (Hagemeijer et al., 2014). More recently, the CREOLEVAL initiative introduced a multilingual POS tagging benchmark covering *Kreyòl*, Mauritian Creole, and over two dozen other Creoles (Lent et al., 2024). While CREOLEVAL marks a significant step forward, its datasets are still modest in size and often restricted to specific domains. *Kreyòl*'s inclusion in CREOLEVAL provides one of the first standardized POS-tagged corpora for the language, albeit relatively small. Many widely spoken Creoles (e.g. Jamaican Creole or Papiamento) still lack any publicly available POS-labeled corpus. This underscores that Creoles remain "*extremely under-explored*" in NLP resource development (Armstrong et al., 2022), with POS annotation efforts only beginning to emerge.

Publicly available treebanks and parsed corpora for Creole languages are even scarcer. Until quite recently, to our knowledge, the only Creole with a Universal Dependencies (UD) treebank was Nigerian Pidgin (Naija, Caron et al., 2019). The UD NAIJA SPOKEN CORPUS consists of over 9,200 tran-

[2] See article by New York Times/Reuters (2003).

[3] Recordings are licensed CC BY-NC-SA 4.0 by the original provider (Duke University, 2024); our annotations are distributed under the same license.

Set	Num. Recordings	Num. Speakers	Raw Duration	VAD Duration	Overlap Duration
Full corpus	3054	1429	1377h:32m	1165h:17m	19h:21m:55s
Sampled (All)	217	223	59h:27m	49h:15m	0h:07m:12s
Sampled (train)	186	195	49h:20m	40h:54m	0h:05m:44s
Sampled (val)	10	20	4h:06m	3h:24m	0h:00m:18s
Sampled (test)	21	36	5h:59m	4h:56m	0h:01m:09s

(a)

Set	Num. Files	Num. Segments	Total Tokens	Avg. Spk/File	Avg. Seg. /Speaker	Avg. Seg. Duration	Avg. Tok/Seg
Full corpus	3054	1,230,670	13,158,884	5.79	80.46	03s	10.69
Sampled (All)	217	46,610	540,072	3.73	68.85	03s	11.59
Sampled (train)	186	38,352	447,058	3.57	69.49	03s	11.66
Sampled (val)	10	2912	37,295	4.70	73.08	04s	12.81
Sampled (test)	21	5346	55,719	4.71	61.14	03s	10.42

(b)

Table 1: Overview of the RADIO HAITI corpus. (1a) Global summary of recordings, speakers, durations, and overlaps by split. *Sampled* correspond to the part of the corpus that will be made fully public. (1b) Transcription statistics by split, including number of files and segments, total tokens, average speakers per file, average segments per speaker, mean segment duration, and mean tokens per segment.

scribed spoken sentences (about $140,000$ tokens) from dialogues and monologues, annotated with dependency syntax. This remains one of the largest syntactic resources for any Creole, and notably a *spoken* one. By contrast, *Kreyòl* obtained its first dependency treebank only in 2023. The UD-HAITIAN-CREOLE-AUTOGRAMM[4] corpus includes 144 sentences (comprising 3.4k tokens) from Bible translations, literature, and news text (Kahane et al., 2024). While valuable as a start, this Haitian treebank's small size and focus on formal written genres limit its coverage of everyday or spoken *Kreyòl*. In the last UD release (v2.16), a major resource for *Kreyòl* was released: the UD-HAITIAN-CREOLE-ADOLPHE[5] and consists of $3,314$ sentences and over 300k tokens derived from Bible-related sources. Other Creoles lag behind. In some cases, researchers have attempted to leverage a Creole's lexifier language to bootstrap parsing: for Martinican Creole, Mompelat et al. (2022) built a dependency parser by transferring from French treebanks. That approach yielded a preliminary treebank (236 sentences) but also exposed mismatches in morphosyntax between this Creole and French, leading to tagging inconsistencies. These results underscore the limitations of cross-lingual parsing without Creole-specific training data. Syntactically annotated corpora for Creole languages remain limited to small projects, but inclusion of Creole into UD and multilingual benchmarks, signal progress in addressing this gap.

Comparable Initiatives for French. Regarding French, the ESLO corpora were collected in the 1970s and 2010s to study sociolinguistic variations in the Orléans region. Abouda et al. (2006) report that the combined corpora consist of more than 700h of speech, comprising roughly 10M words. Even if there seems to have been some work regarding POS tagging (Eshkol et al., 2010), these annotations have not been released. More recently, the ORFÉO treebank was released, (Benzitoun et al., 2016) and contains 164h of spontaneous spoken French (3.5M tokens), annotated in dependency syntax and parts of speech. Approximately 5% of ORFÉO has been annotated manually while the rest was automatically tagged and parsed.

3. Data & Annotation

In this section, we describe the data and its origin, the annotation procedure, and the tools used to produce the corpus, from ASR and its derived annotations (alignments, confidence scores) to POS tagging.

3.1. Source

The corpus we introduce is derived from recordings of the RADIO HAÏTI-INTER radio station. Originally established as RADIO HAÏTI in the late 1930s, the station was renamed RADIO HAÏTI-INTER in the 1970s following its acquisition by the journalist Jean Léopold Dominique, who resumed its direction, until his assassination in the early 2000s, after which the radio stopped operating. The station was at the forefront of the defense of democracy,[6] and consequently, a significant portion of the recordings addresses political matters. Throughout its history, in addition to political matters, RADIO HAÏTI-INTER broadcast daily programs addressing a broad spectrum of topics, including news, roundtable discussions, speeches, and editorials. The archive encompasses both in-studio productions and field recordings, notably naturalistic street interviews.

The recordings were donated to Duke University and have been made publicly accessible through the *David M. Rubenstein Rare Book & Manuscript Library* (Duke University, 2024). The full dataset consists of more than $5,000$ recordings, corresponding to approximately $2,400$ hours of spoken content. Roughly half of these recordings are exclusively in *Kreyòl*, while the remainder primarily feature French, or a combination of French and *Kreyòl*, with limited occurrences of English and Spanish. For the purposes of the present work, we restrict

[4] https://github.com/UniversalDependencies/UD_Haitian_Creole-Autogramm (last accessed: 2025-10-16).

[5] https://github.com/UniversalDependencies/UD_Haitian_Creole-Adolphe (last accessed: 2025-10-16)

[6] See Reporters Without Borders (2003).

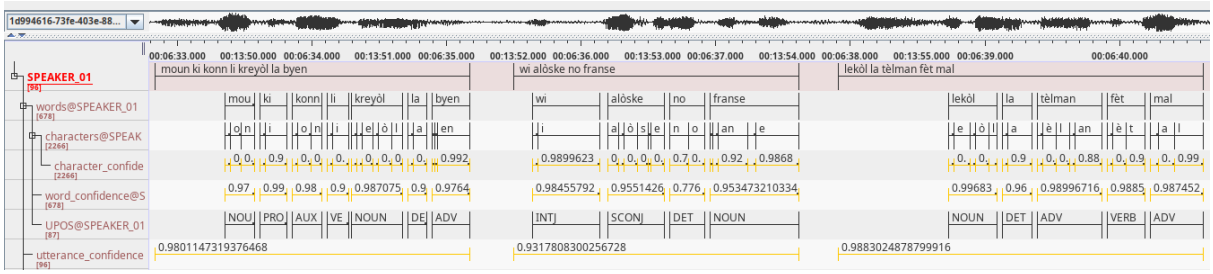

Figure 2: Example alignment viewed in ELAN. Sections that are incorrectly transcribed (e.g. *no* instead of *nan*) receive a lower confidence score. An EAF file can be generated for each recording from the CSV and JSON outputs containing the alignments. Confidence scores at the utterance, word, and character levels are displayed as yellow bars in the tiers. Note that word boundaries are marked by space characters, which leads to unreliable alignments at these positions. In contrast, intra-word alignments and non-boundary segments are generally stable and accurate.

our analysis to the *Kreyòl* subset of the corpus, comprising approximately $2,900$ recordings and exceeding $1,300$ hours of audio material.

3.2. Automatic Speech Recognition

We used one of the models introduced by Havard et al. (2025) to transcribe all the recordings. These models were pre-trained from scratch and have only seen *Kreyòl* (i.e. they are not based on any multilingual model). Once pre-trained, they were then fine-tuned on a speech recognition task.

The models were trained on a wide pool of corpora, consisting of more than 2000h of raw speech captured in a wide range of environmental conditions. The datasets include read speech from HAITI-CMU (Carnegie Mellon University, 2010), as well as spontaneous speech captured over the phone from IARPA-BABEL HAITIAN CREOLE (Andrus, Tony et al., 2017), but also archival fieldwork recordings material from the ATLAS LINGUISTIQUE D'HAÏTI (FLA et al., 2015), and contemporaneous fieldwork recordings featuring dialectal variation from the CORPUS OF NORTHERN HAITIAN CREOLE (Valdman, 2022). The training dataset also includes contemporaneous material from VOXLINGUA HAITIAN (Valk et al., 2021) scrapped from YouTube. Finally, the models were also pre-trained on RADIO HAÏTI-INTER itself, which represented roughly 70% of the training material.

All the spoken material was pre-processed by Havard et al. (2025) using a standardised pipeline: the recordings were resampled to 16kHz 16-bits mono PCM files. A Voice Activity Detection (VAD) model and diariser (Bredin et al., 2020) was applied to detect spoken sections and remove non-spoken sections and noise from the recordings, and automatically assign the spoken sections to speakers.

Once pre-trained, Havard et al. (2025) fine-tuned the models on a speech recognition task using HAITI-CMU and IARPA-BABEL HAITIAN CREOLE with CTC loss (Graves et al., 2006). Havard et al. (2025) made sure to normalise the text to only contain graphemes officially recognised by the Haitian Creole Academy (*Akademi* Kreyòl *Ayisyen*).[7]

In this work, we use the DATA2VEC model they introduce, which is reported by Havard et al. (2025) to be the best performing model with a Word Error Rate (WER) of 33.7% and a Character Error Rate (CER) of 21.0%. As a matter of comparison, an XLSR-2-300M-LARGE (Babu et al., 2022) model fine-tuned on the same set obtains a WER of 35.5% and CER of 21.5%. The authors report that fine-tuned MMS-based models (MMS-1B-ALL, MMS-1B-FL102, Pratap et al. 2024) obtain even higher error rates, with WERs as high as 60% and CERs of more than 30%. Hence, the models introduced by Havard et al. (2025), and more specifically the DATA2VEC model, are up-to-now, the best available models to transcribe *Kreyòl*.

3.3. Alignments

During decoding (i.e. automatic speech transcription), we preserve the alignment between each frame and the character predicted by the model. This yields a direct mapping between the raw speech signal and the predicted character sequence. Words are reconstructed by defining a word as the sequence of characters occurring between two space symbols (see Figure 2 for an illustration). Because the CTC model explicitly predicts spaces to delimit words, the alignments at word boundaries are necessarily unreliable: a portion of the acoustic signal corresponding to either the preceding or following phoneme is aligned to a space character, resulting in unfaithful boundary alignments. Apart from this edge case at word boundaries, the frame-to-character alignments within words are stable and accurate.

Some recordings, however, contain background music, sung speech, speaker overlaps, or various

[7] $\mathcal{V}$ = a à b c d e è f g h i j k l m n o ò p r s t u v w y z | ; where | represents the space character.

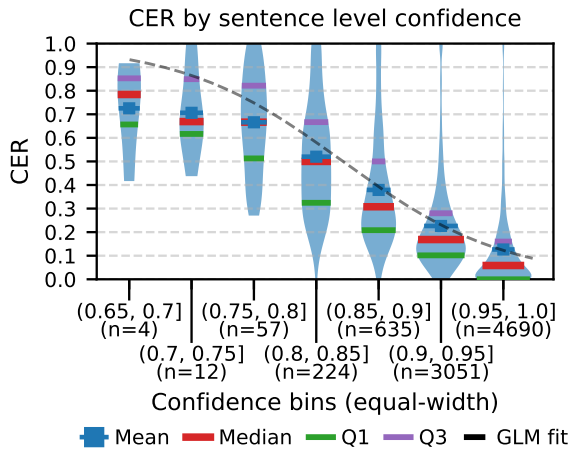

Figure 3: Violin plots showing the distribution of CER by binned sentence-level confidence. Overall, the higher the confidence, the lower the CER. n signals the number of samples in each bin.

environmental noises. In these situations, the alignment becomes unreliable due to the model's limited ability to transcribe the signal faithfully. To detect such cases, we introduce confidence scores (subsection 3.4). All in all, the alignments we provide will enable linguists to focus on specific portions of the spoken input (e.g. particular words or sounds), which can be especially useful for phonetic and phonological studies.

3.4. Confidence Measure

Computation. As our primary audience consists of linguists, we augment the transcriptions with confidence scores. For linguistic analysis, it is essential not only to provide transcriptions, but also to indicate how *reliable* these transcriptions are, which are reflected by the uncertainty of ASR model. Since gold annotations are not available for the entire dataset, and inspired by Laptev et al. (2022), we add entropy-based confidence scores for each utterance, word, and character.

At each timestep t, the CTC model outputs a logit vector $\mathbf{z}_t$ over the vocabulary of graphemes $\mathcal{V}$.[7] The probability of each item of the vocabulary $p_t(v)$ is given by a $\mathrm{Softmax}$ as in Equation 1. Using these values, we are able to compute the entropy H for frame t (Equation 2).

$$p_t(v) = \frac{\exp(z_{t,v})}{\sum_{v' \in \mathcal{V}} \exp(z_{t,v'})} \quad (1) \qquad H_t = -\sum_{v \in \mathcal{V}} p_t(v) \log p_t(v). \quad (2)$$

For each span S_i of length L_i frames (e.g. corresponding to an utterance, word, or character, using character or word alignments described above), we compute the average entropy $\bar{H}_i$ as in Equation 3, where the sum runs over all frames in span i. We then normalise this value to $[0, 1]$ to obtain the normalised per-span entropy $\tilde{H}_i$ (Equation 4), where $V = |\mathcal{V}|$ is the vocabulary size, and finally define the per-span confidence score C_i using Equation 5.

$$\bar{H}_i = \frac{1}{L_i} \sum_{t \in S_i} H_t \quad (3) \qquad \tilde{H}_i = \frac{\bar{H}_i}{\log V} \quad (4) \qquad C_i = 1 - \tilde{H}_i \quad (5)$$

Evaluation. We used the test split used by Havard et al. (2025) to assess the relationship between sentence-level confidence and average character error rate per sentence (see Figure 3). Higher confidence was strongly associated with lower error rates. One-sided correlation tests yielded Pearson $r = -0.48$, Spearman $\rho = -0.57$, and Kendall $\tau = -0.42$ (all $p < 10^{-16}$). A fractional-logit binomial GLM confirmed this pattern: confidence has a large negative effect on error rate ($\beta = -15.29$, SE $= 0.36$, $z = -42.73$, $p < 10^{-16}$; 95% CI $[-15.99, -14.59]$), implying that a 0.10 increase in confidence corresponds to an $\approx 78\%$ reduction in the odds of error.

3.5. POS Tagging

Computation. We automatically annotated the transcriptions with POS tags using the Universal Dependencies (UD) tagset (UPOS). To do so, we trained a custom *Kreyòl* tagger backed by the BERT-based (Devlin et al., 2019) neural parser implementation of Guiller (2020). We fine-tuned the model to predict UPOS tags on this data, using the larger UD-Haitian-Creole-Adolphe for training and the UD-Haitian-Creole-Autogramm corpus for development, using XLM-RoBERTa-large embeddings (Conneau et al., 2020).

We then adapted the initial tagger on one full Radio Haïti-Inter recording (not part of the 50h release) consisting of 394 sentences ($7{,}778$ tokens) for 5 speakers. We first corrected ~ 2.5k tokens according to UD guideline (with adaptations for spoken *Kreyòl*) and use this portion as a development set to tune hyperparameters. Our best setup reached 81.3% UPOS accuracy.[8] We then used this model to pre-label the remaining data, which substantially sped up manual annotation of the full 7.8k-token sample.

Finally, we retrained the model on our newly annotated sample to better adapt it to the spoken genre. The final model was then used to automatically tag all transcripts in the corpus and achieves an overall tagging accuracy of 88.7%, demonstrating a substantial improvement over the baseline .

[8] Full experimental results will be provided in the appendix of the camera-ready version.

Evaluation. Performance varies across POS categories.[8] Among the most frequent categories, content words such as NOUN ($F_1 = 0.90$), VERB (0.91) and PRON (0.94) are identified with high reliability. Functional elements like ADP (0.88), DET (0.89) and AUX (0.95) also reach strong precision and recall values, showing that the model captures major grammatical patterns of *Kreyòl* well. Some classes remain more challenging. Adjectives (ADJ, $F1 = 0.75$), subordinating conjunctions (SCONJ, 0.77) and numerals (NUM, 0.80) tend to be confused with other syntactic categories depending on their position and prosodic context. The lowest scores are observed for X (0.40), which are mainly used to classify disfluencies, such as false starts, underrepresented in the training data. Additionally, we observed difficulties with out-of-vocabulary items such as code-switching instances or French loanwords — which were absent from the UD training data — as well as inconsistencies in spelling of spoken utterances (e.g. phonetic or contracted variants). This phenomena can be extended to proper nouns which suffer from lower recall (PROPN, 0.69).

Despite these challenges, the automatic POS annotations are of sufficient quality for some downstream linguistic and computational analyses. They constitute the first large-scale syntactic layer available for *Kreyòl* speech and establish a strong foundation for further experiments.

4. Human Evaluation

In this section we describe the manual evaluation of the quality of the automatic annotations.

Transcriptions. We recruited an external expert annotator to assess transcription quality. Due to the scarcity of specialists in French-based Antillean Creoles, we were only able to recruit one annotator. The annotator holds a B.A. in Creole Studies and is a native speaker of Martinican Creole. While Haitian and Martinican Creole are not identical, they are closely related and mutually intelligible. Additionally, the annotator we recruited also had formal training and prior professional experience in *Kreyòl*.

A custom Praat script was created to display the item to evaluate to the annotator. The annotator was asked to annotate whether the recordings contained music (YES/NO), sung speech (YES/NO), and to infer the gender/sex of the speaker (FEMALE/MALE). For all three tasks we included an N/A option if the annotator felt unsure. Finally, the annotator was asked to correct the transcriptions. The annotator was allowed to freely go back and forth between the items to annotate and change their minds. They were asked to skip sections that contained only music if there was not any speech, and to correct the transcription if speech was noticeable. The annotations to evaluate were sampled from the test set that is part of the 50 hours that will be shared. We sampled multiple utterances from the recordings and presented them to the annotator in temporal order to preserve context during evaluation. In order to measure how time consuming this task was, the annotation time for each item was recorded as well as the number of time each item was seen. We limited the evaluation session to two hours.

The global results are presented in Table 2a. All in all, the annotator managed annotating 155 items. The final Word/Character Error Rate (WER/CER) are of 16% and 7% respectively, confirming the high quality of the transcription of the sampled items.

We excluded the first two and the final annotated items from the time and views analyses because they correspond to training (first two) and post-task recollection (final) items. The resulting analytic sample comprised 152 items. On average, each file took 44s to be fully annotated (included transcription revision) and was viewed 1.7 times, indicating some degree of correction and verification. Annotating the 152 items took a total of 1 hour and 52 minutes. The low amount of time (44s) required to correct the transcription reflect that most of the transcriptions didn't require any correction. Indeed, when we focus on the amount of time required to edit the transcriptions when they required some correction, the amount of time jumps to 01m22s (± 1m57s).

Metric	Value
Mean WER (↓)	16%
Mean CER (↓)	7%
Mean time per annotation (all)	44s
Mean time per annotation (WER/CER>0)	01m:22s
Mean views per annotation	1.717
Total task duration	1h:52m
Number of annotations	155

(a)

Sex/Gender	Count
Female	6
Male	149
N/A	0

(b)

	Yes	No	N/A
Music	4	151	0
Sung	0	154	1

(c)

Table 2: (a) Error rates and task metadata; (b) gender/sex; (c) music and sung speech.

Regarding gender/sex distribution and the amount of music and sung speech, the results are presented in Table 2b and Table 2c. We observe a large gender/sex imbalance with 149 segments attributed to males, and only 6 attributed to females. Regarding the amount of music, 4 segments were reported as containing background music, and no sung speech was reported.

From a qualitative point of view, the annotator mentioned that they felt the transcription were of high quality, and the majority of the mistakes occurred when proper nouns were used. Additionally they mentioned that the model tended to transcribe "*gen yon*" (to have a) as "*gon*" which is a contraction commonly used in spoken *Kreyòl*, reflecting a potential bias in the transcriptions of the training data use by Havard et al. (2025).

POS Tagging. Following the manual revision of transcriptions described in the previous section, the same 155 sentences were used to conduct a human validation of the POS tagging. This validation was performed on the predictions produced by the model introduced in subsection 3.5, which was trained on UD *Kreyòl* treebanks and adapted to our corpus with a gold sample.

For each sentence, the automatic POS tags were compared against the manually corrected version. This evaluation allows us to estimate the reliability of the automatic annotations in realistic speech conditions, since the material includes spontaneous speech, disfluencies, and orthographic inconsistencies that reflect pronunciation variants. The global accuracy measured on this evaluation set reaches 82.5%. These results confirm that the automatic tagging achieves a level of precision consistent with other low-resource spoken corpora while maintaining robustness across diverse speakers, registers and topics.

POS	Precision	Recall	F1	# tokens
ADJ	0.691	0.595	0.639	79
ADP	0.624	0.852	0.721	115
ADV	0.816	0.775	0.795	120
AUX	0.955	0.970	0.962	132
CCONJ	0.953	0.788	0.863	52
DET	0.863	0.819	0.841	177
INTJ	0.182	0.286	0.222	7
NOUN	0.821	0.871	0.845	363
NUM	0.639	0.622	0.630	37
PRON	0.920	0.941	0.931	256
PROPN	0.421	0.178	0.250	45
SCONJ	0.932	0.585	0.719	94
VERB	0.858	0.912	0.884	251
X	0.000	0.000	0.000	0
Accuracy (↑)		0.825		

Table 3: Detailed evaluation of the POS tagger on the manually validated subset ($n = 155$ sentences).

A closer inspection of the per-category results in Table 3 reveals decent performances for the major grammatical classes of *Kreyòl*. Auxiliary verbs (AUX), pronouns (PRON), determiners (DET), verbs (VERB) and nouns (NOUN) obtain the highest scores, with F_1 values ranging between 0.84 and 0.96. These categories account for most tokens in the corpus and form the backbone of *Kreyòl*'s morphosyntax, which explains the model's stability on these classes. The model also performs well on adverbs (ADV) and conjunctions (CCONJ, SCONJ) indicating that sentence-internal functional patterns are captured reliably. However, performance decreases for certain categories characterised by high contextual or orthographic variability, such as adjectives (ADJ), numerals (NUM), and proper nouns (PROPN). For these tags, F_1 scores range from 0.25 to 0.64, reflecting confusions with other syntactic categories or inconsistent casing.

Overall, the human validation confirms that the model's predictions are mainly consistent with expert annotations for the structural core of the language. The main sources of error stem from lexical sparsity, occasional orthographic inconsistencies in transcriptions, and the fluid utterances boundaries of spoken data. The performance also depends on the segmentation of the data into utterances, which was not manually controlled and relies on the consistency of the VAD model (Bredin et al., 2020). At present, the classifier is trained using the maximal analysis unit, but future work will focus on developing a more suitable strategy for highly variable segmentation of spoken data in POS classification.

Despite these limitations, the model proves robust enough to provide large-scale morphosyntactic annotation of our data, establishing a foundation for subsequent syntactic and diachronic analyses of *Kreyòl*.

5. Radio Haïti-Inter Corpus

In this section we present statistics on the whole corpus and the controlled part that will be publicly released.

Constitution. The statistics for the full corpus and the sampled part are shown in Table 1. In terms of duration, the whole corpus comprises 1377 hours of raw speech, reduced to 1165 hours once VAD was applied. Overall, the amount of overlap between speakers is rather small (20h of overlapping speech across the 1377h of speech), mainly due to the nature of the source material (radio broadcasts). On average the diariser detected 5.7 speakers per file and created 1.2M segments, which are rather short (3 seconds on average).

The part of the corpus that will be freely distributed consists of 50h of spoken material that were automatically transcribed (see subsection 3.2), sampled from the whole corpus. The sampled recordings were chosen to maximise overall quality: sampling was based on the quality of the produced transcriptions, as assessed by confidence scores (see subsection 3.4). Additionally, selection was made to minimise speaker overlap, while ensuring that the sampled recordings span a substantial

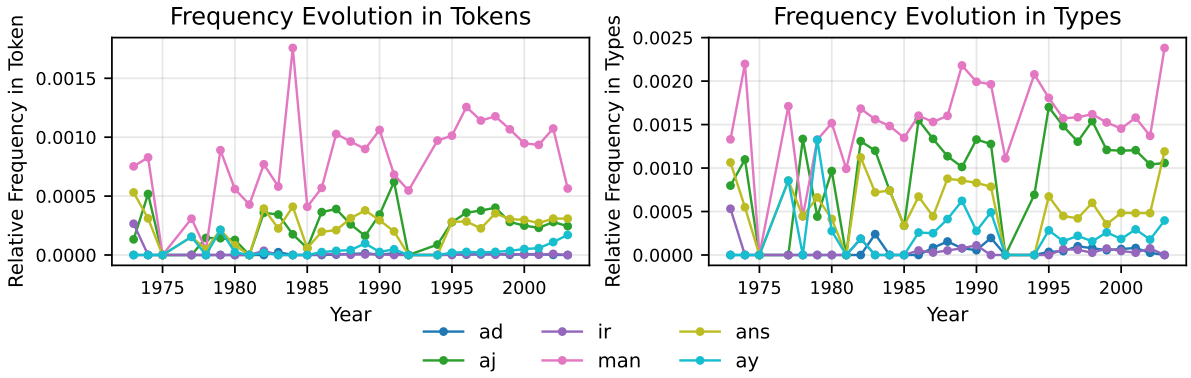

Figure 4: Relative frequency evolution of the Event-denoting suffixes, in Tokens (left) and Types (right)

range of recording years (see Figure 1)

As we believe this portion of the corpus will be used by language technologists, we ensured that it reflects the original train/val/test split of Havard et al. (2025) used to train the transcription models. This will allow language technologists to train and test models in controlled settings, by ensuring that there is no content leakage across splits and that evaluation remains comparable and unbiased across experiments. Respecting the original splits, along with confidence score, also enables guiding the annotation efforts by prioritising which transcriptions should be revised for each split.

Formats & Distribution. The corpus is distributed in a structured and interoperable format to facilitate both linguistic and computational use. The root directory contains four main folders: (*i*) a folder with the original audio recordings in `.wav` format; (*ii*) a folder with time-aligned transcriptions in ELAN's `.eaf` format, compatible with multimodal annotation workflows (Brugman et al., 2004); (*iii*) a folder with plain-text transcriptions in tabular `.csv` and `.json` formats, enabling easy import into NLP pipelines; and (*iv*) a folder containing metadata files describing each recording, including speaker information, recording year, and duration statistics. Each of these folders is further divided into three subfolders corresponding to the data splits — `train/`, `dev/`, and `test/`.

Use Cases. As previously mentioned, the corpus is suited for a wide variety of linguist analyses, from phonetics to semantics, as well as the study of code-switching and patterns of disfluencies (among many other topics). We demonstrate here a short use case, focusing on morphology, and the study of suffixes. There have been long standing debates on whether Creole languages, and in particular *Kreyòl*, have derivational morphology, some (d'Ans, 1968, p. 26) arguing that it was "*logically impossible*" for these languages to have any. Even if this view has been disproved since (DeGraff, 2001; Lefebvre, 2003), the study of the productivity of some affixes has been debated, mainly because of the absence of large scale corpora. Our corpus enables the study of this phenomenon. We present in Figure 4 the distribution of several event-denoting suffixes (-aj/ay, -ad, -man, -ir, -ans) across the years 1973–2003, in both tokens and types. Although "-aj" appears only modestly present when considering token counts, the type counts reveal a broad base coverage, indicating that it attaches to a wide variety of stems. This small case study illustrates how our corpus can be used to investigate affix productivity in *Kreyòl*.

6. Discussion

The corpus we present is only a first step toward enabling the study of spoken *Kreyòl*; but remains a major step, as it is the first time such a large corpus will be made available – regardless of whether we consider the full corpus or the portion that will be made publicly available.

However, our work comes with a few limitations. Although the dataset was manually annotated by a domain expert, all labels were reviewed by a single annotator. Moreover, the annotated subset is restricted to recordings linked to the highest-confidence automatic transcriptions. Consequently, the evaluation is likely biased and reflects only a narrow portion of the data. For example, we observed recordings with music that were transcribed even though the lyrics were not in *Kreyòl*; such cases are not represented in the annotated subset. We believe these segments are rare, but they are likely underrepresented in our evaluation. We make a similar observation regarding gender/sex imbalance: even though there are a majority of male speakers in our dataset, it may not be as biased as what our evaluation reflects.

As future work, we aim to extend the annotations by including not only POS tags but also dependency trees, enabling fine-grained studies of the morphosyntax of spoken *Kreyòl*. We also plan to

annotate code-switched and mixed segments to examine the porosity between *Kreyòl* and its lexifier language, French. Finally, an important direction is to undertake a larger round of manual annotation, as the one we provide, even though carefully curated and validated, remains limited. This would allow us to release a larger portion of the dataset, extending beyond the 50h we release.

7. Conclusion

In this paper, we presented the RADIO HAÏTI-INTER corpus, the first large-scale corpus of spoken *Kreyòl*, built with dedicated state-of-the-art models. It comprises transcriptions, enriched with POS tags, time-aligned transcripts, and confidence scores to facilitate selective, high-reliability analyses. A manual evaluation of the released portion reports a character error rate of 7% and a POS-tagging accuracy of 82.5%, underscoring the corpus's suitability for research. To enable immediate, high-quality use, we release 50 hours of audio with accompanying annotations drawn from the most reliable segments. This resource opens new avenues from phonetics to semantics, and further supports investigations of code-switching, patterns of disfluency, and micro-diachronic variation in *Kreyòl* over the many years spanned by the recordings.

8. Acknowledgements

We thank Duke University and the David M. Rubenstein Rare Book & Manuscript Library for making the original data publicly available through the Radio Haiti Archive (https://repository.duke.edu/dc/radiohaiti). The authors also acknowledge the support of the French Agence Nationale de la Recherche (ANR) under grant ANR-20-CE38-0006 (project CREAM). Experiments were conducted using Grid'5000, developed under INRIA ALADDIN with support from CNRS, RENATER, and several universities. Additional computational resources were provided by the CaSciModOT cluster at the Centre de Calcul Scientifique en région Centre-Val de Loire and by the HPC resources of IDRIS made available through GENCI under allocation 2024-AD011014940. We further gratefully acknowledge the feedback of Pr. Renauld Govain (Faculté de Linguistique Appliquée, Haïti) for regular exchanges and for beta testing the corpus.

9. Bibliographical References

Lotfi Abouda and Olivier Baude. 2006. Constituer et exploiter un grand corpus oral : choix et enjeux théoriques. le cas des eslo. In *Corpus en Lettres et Sciences sociales, Des documents numériques à l'interprétation*, Albi, France.

Ruth-Ann Armstrong, John Hewitt, and Christopher Manning. 2022. JamPatoisNLI: A jamaican patois natural language inference dataset. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5307–5320, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2022. XLS-R: Self-supervised cross-lingual speech representation learning at scale. In *Interspeech 2022*, pages 2278–2282.

Olivier Baude and Céline Dugua. 2016. Les eslo, du portrait sonore au paysage digital. *Corpus*, (15).

Christophe Benzitoun, Jeanne-Marie Debaisieux, and Henri-José Deulofeu. 2016. Le projet ORFÉO: un corpus d'étude pour le français contemporain. *Corpus*, (15).

Hervé Bredin, Ruiqing Yin, Juan Manuel Coria, Gregory Gelly, Pavel Korshunov, Marvin Lavechin, Diego Fustes, Hadrien Titeux, Wassim Bouaziz, and Marie-Philippe Gill. 2020. pyannote.audio: neural building blocks for speaker diarization. In *ICASSP 2020*, Barcelona, Spain.

Hennie Brugman and Albert Russel. 2004. Annotating multi-media/multi-modal resources with ELAN. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Bernard Caron. 2020. Methodological and technical challenges of a corpus-based study of Naija. In Nina Pawlak and Izabela Will, editors, *West African Languages: Linguistic Theory and Communication*. University of Warsaw Press, Warsaw.

Bernard Caron, Marine Courtin, Kim Gerdes, and Sylvain Kahane. 2019. A surface-syntactic ud treebank for Naija. In *Proceedings of the 18th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2019)*, pages 13–24. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020.

Unsupervised cross-lingual representation learning at scale.

André-Marcel d'Ans. 1968. *Le créole français d'Haïti*. Mouton de Gruyter, The Hague.

Michel A. F. DeGraff. 2001. Morphology in creole genesis: A prolegomenon. In Michael Kenstowicz, editor, *Ken Hale: A Life in Language*, pages 53–121. The MIT Press, Cambridge, MA.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig. 2024. *Ethnologue: Languages of the World*, 27 edition. SIL International, Dallas, Texas. Accessed: 2025-10-10.

Iris Eshkol, Isabelle Tellier, Samer Taalab, and Sylvie Billot. 2010. Étiqueter un corpus oral par apprentissage automatique à l'aide de connaissances linguistiques. In *Actes des 10èmes Journées internationales d'Analyse statistique des Données Textuelles (JADT 2010)*, pages 1–12, Rome, Italy.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, page 369–376, New York, NY, USA. Association for Computing Machinery.

Kirian Guiller. 2020. Analyse syntaxique automatique du pidgin-créole du Nigeria à l'aide d'un transformer (BERT): Méthodes et résultats. *Mémoire de Master, Sorbonne Nouvelle*.

Tjerk Hagemeijer, Michel Généreux, Iris Hendrickx, Amália Mendes, Abigail Tiny, and Armando Zamora. 2014. The Gulf of Guinea creole corpora. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 523–529, Reykjavik, Iceland. European Language Resources Association (ELRA).

William N. Havard, Renauld Govain, Benjamin Lecouteux, and Emmanuel Schang. 2025. Self-Supervised Models of Speech Processing for Haitian Creole. In *Interspeech 2025*, pages 4018–4022.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 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*.

Sylvain Kahane, Claudel Pierre-Louis, Sandra Jagodzińska, and Agata Savary. 2024. The first Haitian creole treebank. In *Peer reviewed poster in the 2nd UniDive Workshop*.

Aleksandr Laptev and Boris Ginsburg. 2022. Fast entropy-based methods of word-level confidence estimation for end-to-end automatic speech recognition. *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 152–159.

Claire Lefebvre. 2003. *The emergence of productive morphology in creole languages: the case of Haitian Creole*, pages 35–80. Springer Netherlands, Dordrecht.

Heather Lent, Kelechi Ogueji, Miryam de Lhoneux, Orevaoghene Ahia, and Anders Søgaard. 2022. What a creole wants, what a creole needs. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6439–6449, Marseille, France. European Language Resources Association.

Heather Lent, Kushal Tatariya, Raj Dabre, Yiyi Chen, Marcell Fekete, Esther Ploeger, Li Zhou, Ruth-Ann Armstrong, Abee Eijansantos, Catriona Malau, Hans Erik Heje, Ernests Lavrinovics, Diptesh Kanojia, Paul Belony, Marcel Bollmann, Loïc Grobol, Miryam de Lhoneux, Daniel Hershcovich, Michel DeGraff, Anders Søgaard, and Johannes Bjerva. 2024. CreoleVal: Multilingual multitask benchmarks for creoles. *Transactions of the Association for Computational Linguistics*, 12:950–978.

Ludovic Mompelat, Daniel Dakota, and Sandra Kübler. 2022. How to parse a creole: When Martinican creole meets french. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4397–4406.

New York Times/Reuters. 2003. Haitian radio station to close after renewed threats to staff. *The New York Times*.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. 2024. Scaling speech technology to

1,000+ languages. *Journal of Machine Learning Research*, 25(97):1–52.

Reporters Without Borders. 2003. Radio haïti inter shuts down.

Jörgen Valk and Tanel Alumäe. 2021. VoxLingua107: a dataset for spoken language recognition. In *Proc. IEEE SLT Workshop*.

10. Language Resource References

Andrus, Tony and Bills, Aric and Conners, Thomas and Crabb, Erin Smith and Dubinski, Eyal and Fiscus, Jonathan G. and Gillies, Breanna and Harper, Mary and Hazen, T. J. and Hefright, Brook and Jarrett, Amy and Le, Hanh and Ray, Jessica and Rytting, Anton and Shen, Wade and Silber, Ronnie and Tzoukermann, Evelyne and Bishop, Judith. 2017. *IARPA Babel Haitian Creole Language Pack IARPA-babel201b-v0.2b*. Linguistic Data Consortium.

Carnegie Mellon University. 2010. *Public Release of Haitian Creole Language Data*.

Duke University. 2024. *Radio Haiti Collection, David M. Rubenstein Rare Book & Manuscript Library, Duke University*. Digital items: 5,314; Total components: 3660; Last Indexed: 2024-12-05.

FLA and Fattier, Dominique. 2015. *Atlas linguistique d'Haïti*. Université de Cergy-Pontoise.

Tonjes Veenstra. 2024. *Spoken Morisien Corpus*. PARADISEC.

Valdman, Albert. 2022. *Corpus of Northern Haitian Creole*. Internet Archive.

Zribi-Hertz, Anne and Schang, Emmanuel and Glaude, Herby. 2012. *CREOLORAL: Présentation du corpus (créole martiniquais et guadeloupéen)*. SFL (Structure Formelle du Langage); LLL (Laboratoire Ligérien de Linguistique).