Speech Technologies with Fieldwork Recordings: the Case of Haitian Creole

William N. Havard[1,2], Renauld Govain[3], Benjamin Lecouteux[2], Emmanuel Schang[1]

[1] LLL, Université d'Orléans, CNRS, 45000 Orléans, France
[2] LIG, Université Grenoble Alpes, CNRS, Grenoble INP, 38000 Grenoble, France
[3] LangSé, Université d'État d'Haïti, Port-au-Prince, Haïti

Honolulu, 4-5 March 2025

- PROJET FINANCÉ PAR L'ANR - PROJECT FUNDED BY THE ANR

Introduction

Introduction

Haitian Creole, a low-resourced language?

- **Low-resourced languages**

Introduction

Haitian Creole, a low-resourced language?

- **Low-resourced languages**
  - mainly from the **perspective of computer scientists**

Introduction

Haitian Creole, a low-resourced language?

- **Low-resourced languages**
  - mainly from the **perspective of computer scientists**
  - many resources collected by linguists and the community (Bird 2020)

Introduction

Haitian Creole, a low-resourced language?

- **Low-resourced languages**
  - mainly from the **perspective of computer scientists**
  - many resources collected by linguists and the community (Bird 2020)
  - often not in a digital format but **existent nonetheless**

Introduction

Haitian Creole, a low-resourced language?

- **Low-resourced languages**
  - mainly from the **perspective of computer scientists**
  - many resources collected by linguists and the community (Bird 2020)
  - often not in a digital format but **existent nonetheless**
- **Legacy Material**
  - archive data (e.g. stored in labs, libraries, etc.)

Introduction

Haitian Creole, a low-resourced language?

- **Low-resourced languages**
  - mainly from the **perspective of computer scientists**
  - many resources collected by linguists and the community (Bird 2020)
  - often not in a digital format but **existent nonetheless**
- **Legacy Material**
  - archive data (e.g. stored in labs, libraries, etc.)
  - collected many years ago (40+ years, could be even more)

Introduction

Haitian Creole, a low-resourced language?

- **Low-resourced languages**
  - mainly from the **perspective of computer scientists**
  - many resources collected by linguists and the community (Bird 2020)
  - often not in a digital format but **existent nonetheless**
- **Legacy Material**
  - archive data (e.g. stored in labs, libraries, etc.)
  - collected many years ago (40+ years, could be even more)

$\longrightarrow$ Can we train state-of-the-art speech processing models using only **fieldwork legacy material**?

Introduction

Haitian Creole, a low-resourced language?

- **Low-resourced languages**
  - mainly from the **perspective of computer scientists**
  - many resources collected by linguists and the community (Bird 2020)
  - often not in a digital format but **existent nonetheless**
- **Legacy Material**
  - archive data (e.g. stored in labs, libraries, etc.)
  - collected many years ago (40+ years, could be even more)

$\longrightarrow$ Can we train state-of-the-art speech processing models using only **fieldwork legacy material**?

$\longrightarrow$ **"mobilise the archive"** approach (Bird 2020)

Fieldwork Data

- Fieldwork Data
  - data collected for **linguistic purposes**

Fieldwork Data

- Fieldwork Data
  - data collected for **linguistic purposes**
  - **not intended for computational applications**

Fieldwork Data

- Fieldwork Data
  - data collected for **linguistic purposes**
  - **not intended for computational applications**
- Challenges
  - analog tape-recorded material $\longrightarrow$ **digitisation**

Fieldwork Data

- Fieldwork Data
  - data collected for **linguistic purposes**
  - **not intended for computational applications**
- Challenges
  - analog tape-recorded material $\longrightarrow$ **digitisation**
  - inherently **noisy** (reverberation, environmental noise)

Fieldwork Data

- Fieldwork Data
  - data collected for **linguistic purposes**
  - **not intended for computational applications**
- Challenges
  - analog tape-recorded material $\longrightarrow$ **digitisation**
  - inherently **noisy** (reverberation, environmental noise)

$\longrightarrow$ Represents the **majority of data available** for most languages

Haitian Creole

- **13M speakers** (Simons et al. 2023)
- in Haiti and by the Haitian diaspora

Haitian Creole

- **13M speakers** (Simons et al. 2023)
- in Haiti and by the Haitian diaspora
- French-based Creole
  - French is called its lexifier language
  - gave Haitian Creole **most of its vocabulary** (Hazael-Massieux 2012)

Related Works

Speech Processing for Creole Languages

- Relatively sparse field, except for notable works:
  - **Haitian Creole**: Breiter 2014
  - **Guadeloupean** and **Mauritian Creole**: Macaire et al. 2022; Le Ferrand et al. 2023; Le Ferrand et al. 2024
  - **Mauritian Creole** (medical domain): Gooda Sahib-Kaudeer et al. 2019

Speech Processing for Creole Languages

- Relatively sparse field, except for notable works:
  - **Haitian Creole**: Breiter 2014
  - **Guadeloupean** and **Mauritian Creole**: Macaire et al. 2022; Le Ferrand et al. 2023; Le Ferrand et al. 2024
  - **Mauritian Creole** (medical domain): Gooda Sahib-Kaudeer et al. 2019

$\longrightarrow$ Speech processing for Creole languages remains **largely unexplored**

$\longrightarrow$ **No open scientifically reproducable models available**

$\longrightarrow$ **No monolingual models available**

Research Questions

- Assumption: Existence of (potentially old) **fieldwork data**
- Real-world use cases: **Field linguists or community of speakers**

Research Questions

- Assumption: Existence of (potentially old) **fieldwork data**
- Real-world use cases: **Field linguists or community of speakers**

- **Research Questions**
  a. Can noisy fieldwork data be used to train SSL models (e.g. WAV2VEC2)?
  b. Train models from scratch or use transfer learning and CPT?
  c. How much data is needed for ASR fine-tuning?
  d. Can models be trained 'on a budget' (1 GPU)?
  e. Influence of lexifier language (French) on CPT?

Continuous Pre-Training (CPT) Approaches

- CPT: **resume pre-training with new unlabelled data**

Continuous Pre-Training (CPT) Approaches

- CPT: **resume pre-training with new unlabelled data**
- Nowakowski et al. 2023
  - explored **CPT for Ainu** speech recognition using old fieldwork data
  - Used 4 GPUs and XLSR-53 model pre-trained on 56k hours of data
  - Multilingual fine-tuning (English, Japanese, Ainu)

Continuous Pre-Training (CPT) Approaches

- CPT: **resume pre-training with new unlabelled data**
- Nowakowski et al. 2023
  - explored **CPT for Ainu** speech recognition using old fieldwork data
  - Used 4 GPUs and XLSR-53 model pre-trained on 56k hours of data
  - Multilingual fine-tuning (English, Japanese, Ainu)

- **Our approach**

  $\longrightarrow$ stricter use of **fieldwork data at all steps**

  $\longrightarrow$ **linguistically motivated transfer** (from French)

  $\longrightarrow$ comparison with **monolingual model** trained **from scratch**

Data

Atlas Linguistique d'Haiti

- **499 audio recordings** ($\sim$ 45mn each) in Haitian Creole
- Collected by (Fattier 1998) between 1978 and 1987 for a linguistic atlas
  - originally recorded on audio cassettes
  - digitised by the French National Library in 2010
- Features interviews eliciting words or phrases from native collaborators

- **Publicly and legally** available

Corpus of Northern Haitian Creole

- Corpus of Northern Haitian Creole
- 10 recorded interviews conducted by (Valdman et al. 2015) in Cap-Haïtien (9 hours)
- Fully transcribed *but* non-standard, impressionistic transcriptions
  - "*Powoprens*" / "*Potoprens*"
  - "*eskeu*" / "*eske*"

- **Focuses on dialectal variation** with regard to standard Haitian

Other Datasets

- **Haiti-CMU**
  - Read speech (~ 20 hours), mainly from the Bible
- **IARPA-Babel**:
  - 203 hours of conversational and scripted telephone speech

Other Datasets

- **Haiti-CMU**
  - Read speech (~ 20 hours), mainly from the Bible
- **IARPA-Babel**:
  - 203 hours of conversational and scripted telephone speech

$\longrightarrow$ used for **out-of-domain testing**

Experiments

Experimental Settings

- **Low-budget constraints**
  - WAV2VEC2-BASE architecture (and not WAV2VEC2-LARGE)
  - excludes fine-tuning multilingual models like XLSR-53

Experimental Settings

- **Low-budget constraints**
  - WAV2VEC2-BASE architecture (and not WAV2VEC2-LARGE)
  - excludes fine-tuning multilingual models like XLSR-53
- Use ALH corpus for pre-training SSL models

Experimental Settings

- **Low-budget constraints**
  - WAV2VEC2-BASE architecture (and not WAV2VEC2-LARGE)
  - excludes fine-tuning multilingual models like XLSR-53
- Use ALH corpus for pre-training SSL models
  - Voice Activity Detection (VAD) model used to isolate speech sections
  - 356h $\longrightarrow$ **229 hours of spoken sections** (**35% shrinkage**)

Experimental Settings

- **Low-budget constraints**
  - WAV2VEC2-BASE architecture (and not WAV2VEC2-LARGE)
  - excludes fine-tuning multilingual models like XLSR-53
- Use ALH corpus for pre-training SSL models
  - Voice Activity Detection (VAD) model used to isolate speech sections
  - 356h $\longrightarrow$ **229 hours of spoken sections** (**35% shrinkage**)
- Implemented using `fairseq` recipe
- Trained on a single GPU with **gradient accumulation to simulate 16 GPUs**

Model Pretraining and Finetuning

Base Model	Pretraining with ALH

Model Pretraining and Finetuning

	Base Model		Pretraining with ALH
Native from scratch	$\varnothing$	$\longrightarrow$	NATIVE-HAT-SSL

Model Pretraining and Finetuning

	Base Model		Pretraining with ALH
Native from scratch	∅	⟶	NATIVE-HAT-SSL
	FOREIGN-FRA-SSL (Parcollet et al. 2023)	⟶	FOREIGN-FRA-SSL+CPT

Model Pretraining and Finetuning

	Base Model		Pretraining with ALH
Native from scratch	∅	⟶	NATIVE-HAT-SSL
Foreign with CPT	FOREIGN-FRA-SSL (Parcollet et al. 2023)	⟶	FOREIGN-FRA-SSL+CPT
	FOREIGN-ENG-SSL (Baevski et al. 2020)	⟶	FOREIGN-ENG-SSL+CPT

Model Pretraining and Finetuning

	Base Model		Pretraining with ALH	Finetuning with CNHC
Native from scratch	∅	⟶	Native-Hat-SSL	
Foreign with CPT	Foreign-Fra-SSL (Parcollet et al. 2023)	⟶	Foreign-Fra-SSL+Cpt	
	Foreign-Eng-SSL (Baevski et al. 2020)	⟶	Foreign-Eng-SSL+Cpt	

Model Pretraining and Finetuning

	Base Model		Pretraining with ALH		Finetuning with CNHC
Native from scratch	∅	⟶	Native-Hat-SSL	⟶	Native-Hat-SSL+Ft
Foreign with CPT	Foreign-Fra-SSL (Parcollet et al. 2023)	⟶	Foreign-Fra-SSL+Cpt		
	Foreign-Eng-SSL (Baevski et al. 2020)	⟶	Foreign-Eng-SSL+Cpt		

Model Pretraining and Finetuning

	Base Model		Pretraining with ALH		Finetuning with CNHC
Native from scratch	∅	⟶	Native-Hat-SSL	⟶	Native-Hat-SSL+Ft
Foreign with CPT	Foreign-Fra-SSL (Parcollet et al. 2023)	⟶	Foreign-Fra-SSL+Cpt	⟶	Foreign-Fra-SSL+Cpt+Ft
	Foreign-Eng-SSL (Baevski et al. 2020)	⟶	Foreign-Eng-SSL+Cpt	⟶	Foreign-Eng-SSL+Cpt+Ft

Model Pretraining and Finetuning

	Base Model		Pretraining with ALH		Finetuning with CNHC
Native from scratch	∅	⟶	Native-Hat-SSL	⟶	Native-Hat-SSL+Ft
Foreign with CPT	Foreign-Fra-SSL (Parcollet et al. 2023)	⟶	Foreign-Fra-SSL+Cpt	⟶	Foreign-Fra-SSL+Cpt+Ft
	Foreign-Eng-SSL (Baevski et al. 2020)	⟶	Foreign-Eng-SSL+Cpt	⟶	Foreign-Eng-SSL+Cpt+Ft
Foreign without CPT	Foreign-Fra-SSL (Parcollet et al. 2023)				
	Foreign-Eng-SSL (Baevski et al. 2020)				

Model Pretraining and Finetuning

	Base Model	Pretraining with ALH	Finetuning with CNHC
Native from scratch	∅	⟶ Native-Hat-SSL	⟶ Native-Hat-SSL+Ft
Foreign with CPT	Foreign-Fra-SSL (Parcollet et al. 2023)	⟶ Foreign-Fra-SSL+Cpt	⟶ Foreign-Fra-SSL+Cpt+Ft
Foreign with CPT	Foreign-Eng-SSL (Baevski et al. 2020)	⟶ Foreign-Eng-SSL+Cpt	⟶ Foreign-Eng-SSL+Cpt+Ft
Foreign without CPT	Foreign-Fra-SSL (Parcollet et al. 2023)	⟶	Foreign-Fra-SSL+Ft
Foreign without CPT	Foreign-Eng-SSL (Baevski et al. 2020)	⟶	Foreign-Eng-SSL+Ft

Finetuning Size and Parameters

- **Various finetuning sizes** used to understand impact on performance:
  - Max (360 minutes), 320, 160, 80, 40, 20, 10, and 5 minutes
- Fine-tuned for 20k steps with CTC loss
- Parameters frozen for the first 10k steps to prevent overfitting
- Text lower-cased and diacritics removed

Language Models

- Trained 2-to-5-gram Language Models (LMs) using KenLM
  - LMs trained on CNHC transcriptions only
  - Separate LM for each training data size

  $\longrightarrow$ Resulted in 32 different LMs (4 n-gram sizes $\times$ 8 train sizes)
- Used to compare **raw decoding and LM-rescored decodings**

Results

Results

Model Type	**WER** ↓	CER ↓	Train Size	Decoding	Rank
Foreign-Fra-SSL+Cpt+Ft	**36.8**	21.6	320	4-gram	**1**
Native-Hat-SSL+Ft	37.4	**21.5**	360 (max)	3-gram	5
Foreign-Eng-SSL+Cpt+Ft	37.5	22.4	320	4-gram	6
Foreign-Fra-SSL+Ft	42.5	24.5	360 (max)	3-gram	27
Foreign-Eng-SSL+Ft	50.4	29.0	320	3-gram	49

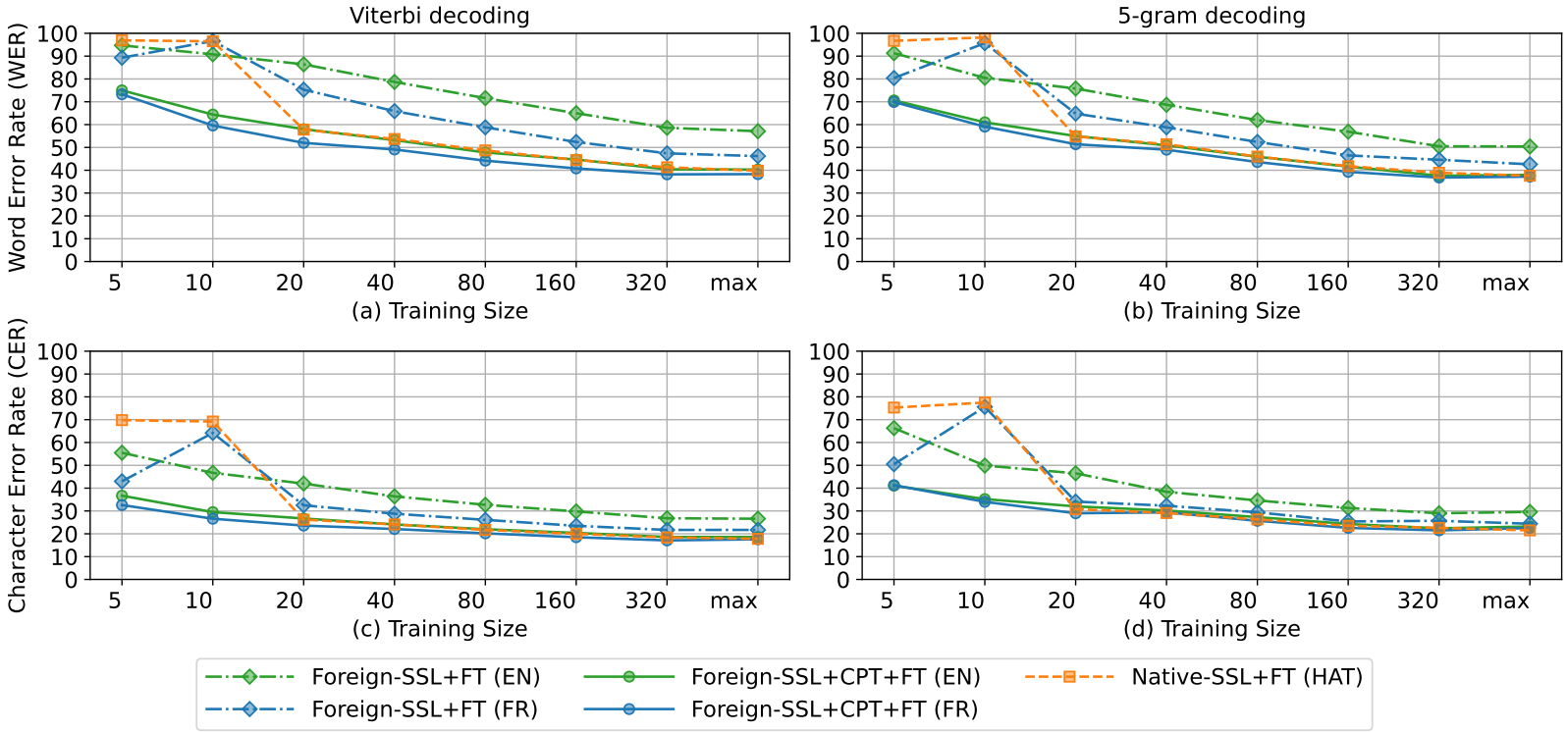
Model Type	WER ↓	**CER** ↓	Train Size	Decoding	Rank
Foreign-Fra-SSL+Cpt+Ft	**38.2**	**17.1**	320	Viterbi	**1**
Native-Hat-SSL+Ft	39.8	17.8	360 (max)	Viterbi	3
Foreign-Eng-SSL+Cpt+Ft	40.3	18.6	360 (max)	Viterbi	6
Foreign-Fra-SSL+Ft	46.2	21.7	360 (max)	Viterbi	12
Foreign-Eng-SSL+Ft	57.1	26.6	360 (max)	Viterbi	38

Fieldwork Data

- Competitive models trained on a **single GPU**
- **Effective use of fieldwork data** for SSL models
- Noisy fieldwork data remains competitive compared to clean data

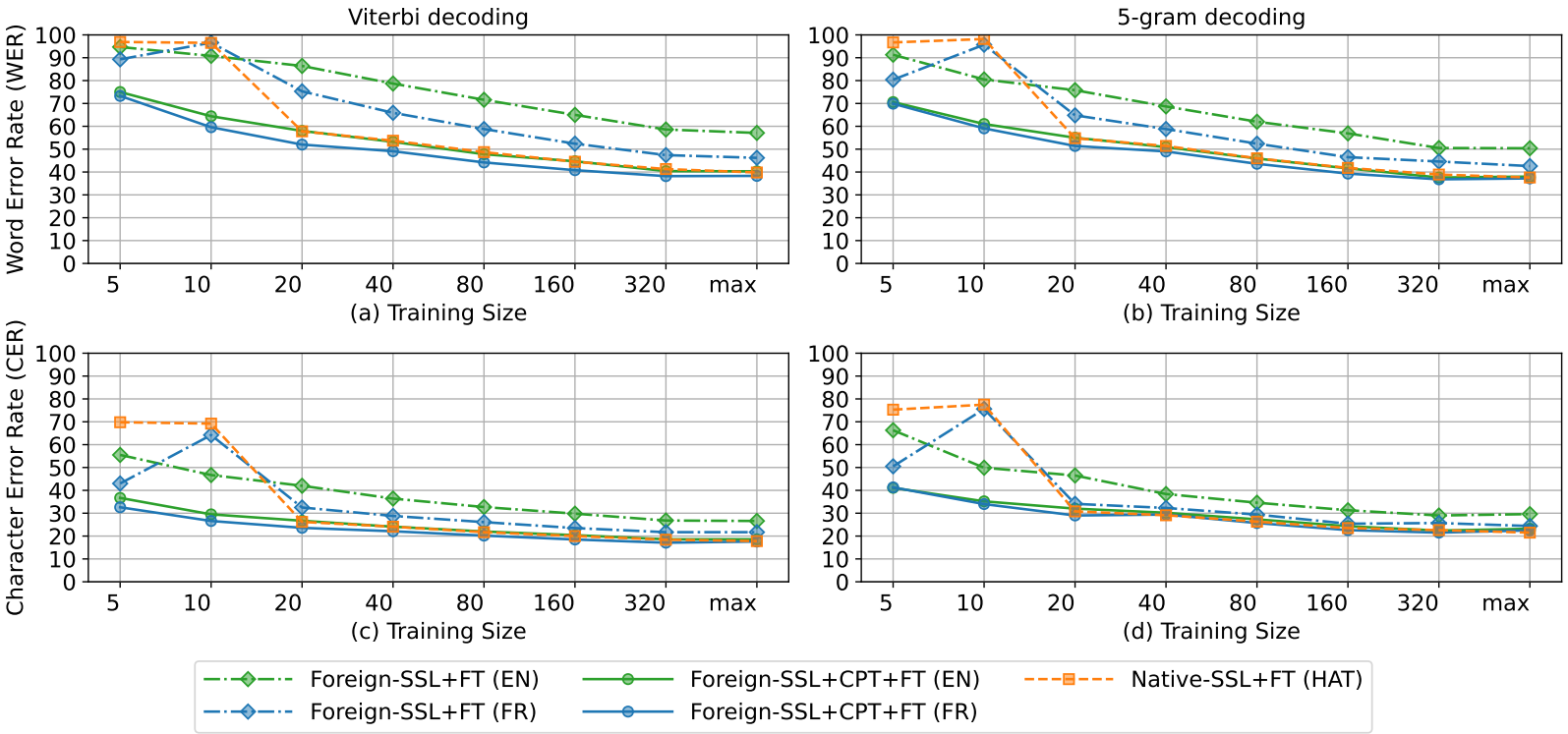
$\longrightarrow$ **Successful Repurposing of old tape-recorded data** for speech processing

Train From Scratch or Use CPT



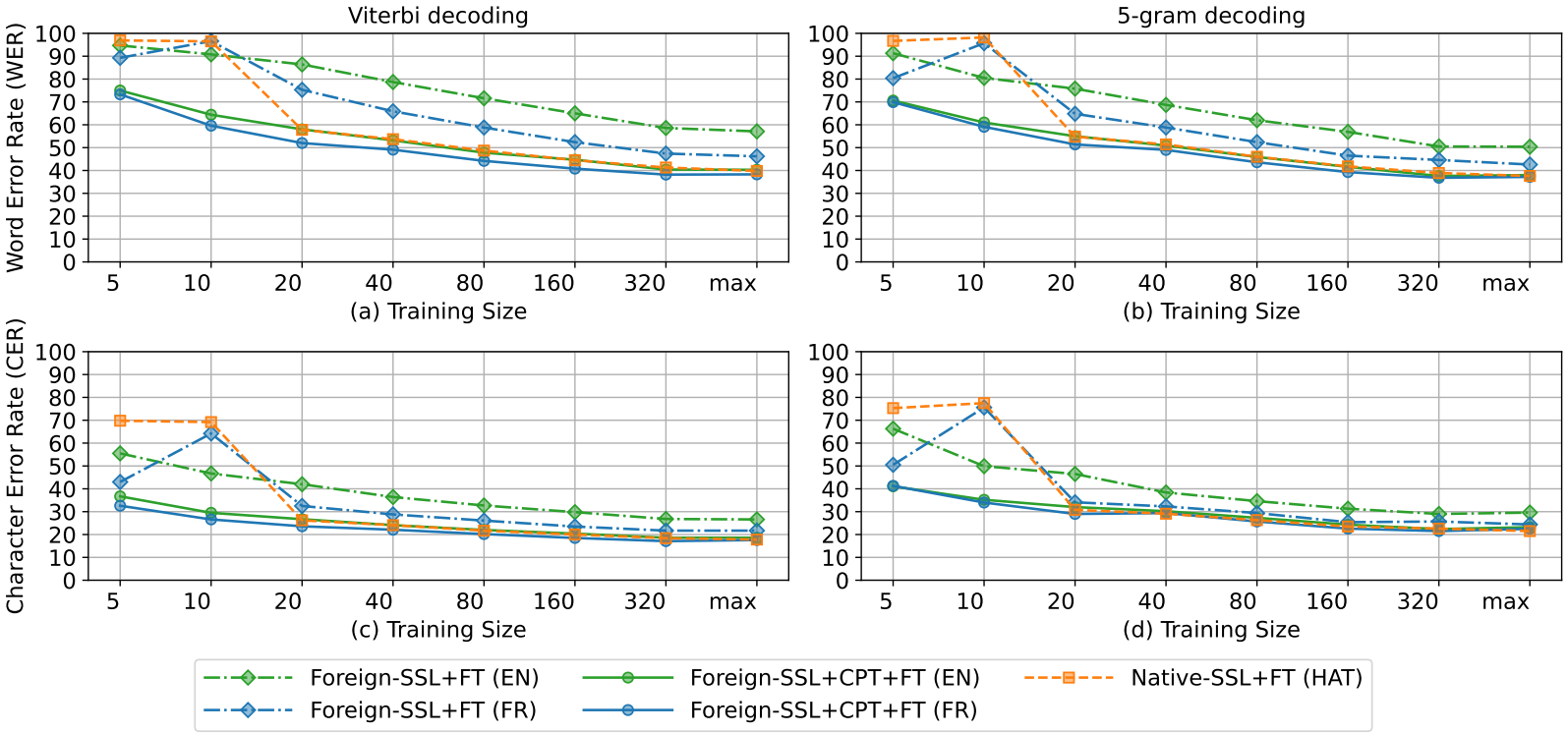
- **Foreign models with CPT show slight advantage** over native models
- Stronger model when CPT is done **on the lexifier language (French)**
- Direct **fine-tuning from SSL models without CPT lags behind**

Amount of Fine-Tuning Data



- Three groups of models with varying robustness to reduced training data
  - Foreign-SSL+Cpt+Ft > Native-Hat-SSL > Foreign-SSL+Ft
- CPT models more robust due to exposure to more speech
- **20 minutes of data closes the gap for the native models**

Viterbi vs. LM Decoding



- **Mixed results with LM decoding**
- LMs significantly **improve WER for directly fine-tuned foreign models**
- LM rescoring **indispensable when no pre-training data** is available

Comparison with MMS

Corpus	Model	CER↓
CNCH	Foreign-Fra-SSL+Cpt+Ft	**17.1**
	Native-Hat-SSL+Ft	17.8
	MMS Pratap et al. 2023	28.4
Haiti-CMU	Foreign-Fra-SSL+Cpt+Ft	09.5
	Native-Hat-SSL+Ft	11.6
	MMS Pratap et al. 2023	**07.9**
IARPA-Babel	Foreign-Fra-SSL+Cpt+Ft	36.6
	Native-Hat-SSL+Ft	38.5
	MMS Pratap et al. 2023	**34.6**

- MMS obtains better scores on the cleaner datasets
- **Our models competitive on fieldwork data**
- Fieldwork recordings **do not hinder zero-shot adaptation**
- **Our models pre-trained on less data** compared to MMS

Future Work

Limitations and Future Work

- Focused on **extrinsic evaluation** of fieldwork data for ASR fine-tuning

$\longrightarrow$ Intrinsic evaluation using ABX task to compare latent representations

Limitations and Future Work

- Focused on **extrinsic evaluation** of fieldwork data for ASR fine-tuning

$\longrightarrow$ Intrinsic evaluation using ABX task to compare latent representations

- Data collected 40 years ago

$\longrightarrow$ impacts of language evolution need to be studied

Limitations and Future Work

- Focused on **extrinsic evaluation** of fieldwork data for ASR fine-tuning

⟶ Intrinsic evaluation using ABX task to compare latent representations

- Data collected 40 years ago

⟶ impacts of language evolution need to be studied

- We used a lot of data ...

⟶ **minimal fieldwork data** required for competitive ASR models (50h pretraining data? 100h?)

Conclusion

Summary of Work

- Used **40-year-old digitised tape-recorded fieldwork data** in Haitian
- Trained a **native SSL model** and used CPT on French and English SSL models
- **Competitive results** achieved, with the best model being the French CPT model

- **Native SSL model also performs well!**

Implications and Contributions

- Demonstrated feasibility of training SSL models using **only fieldwork recordings**
- **Methodology applicable to many languages**, using archived recordings
- **Supports the 'mobilising the archive' approach for speech processing**

We acknowledge the support of the French **Agence Nationale de la Recherche (ANR)**, under grant **ANR-20-CE38-0006 (project CREAM)** coordinated by **Pr. Emmanuel Schang** and **Pr. Benjamin Lecouteux**.

Experiments presented in this paper were carried out using the **Grid'5000 experimental testbed**, being developed under the INRIA ALADDIN development action with support from CNRS, RENATER and several Universities as well as other funding bodies (see `https://www.grid5000.fr`).

We also benefitted from the use of the **CaSciModOT** (`https://cascimodot.fr/`) cluster at the Centre de Calcul Scientifique en région Centre-Val de Loire.

Laboratoire
Ligérien de
Linguistique