

December 6, 2024

Increasing Language Diversity in NLP: Insights from CreoleVal

Marcel Bollmann

 marcel.bollmann.me

 marcel.bollmann@liu.se

NLP Group, Department of Computer and Information Science (IDA)



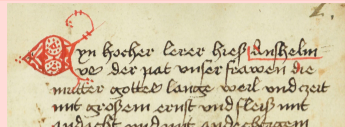
Most slides were adapted from Heather Lent.

- 1 PhD in Computational Linguistics,
Ruhr-Universität **Bochum** 🇩🇪 2012–2018
- 2 Postdoc in CoAStal NLP Group,
University of **Copenhagen** 🇩🇰 2018–2021
- 3 Assistant Professor in Jönköping AI Lab,
Jönköping University 🇸🇪 2021–2023
- 4 Associate Professor in LiU NLP Group,
Linköping University 🇸🇪 since 2023



Historical documents

- Dealing with spelling variation
- Making old texts accessible for research



Multilinguality & Lesser-resources languages



- Transfer learning to LRLs
- **Creating datasets** for Creole languages

TrustLLM: Trustworthy large language models for Europe

- Efficiently adapting LLMs to new languages
- Improving LLMs on “smaller” Germanic languages



Outline

1. NLP for Creole Languages

- What are Creoles?
- Why work on Creoles?
- What resources exist?

2. CreoleVal

- Overview
- Machine Comprehension
- Relation Classification
- Things we didn't do

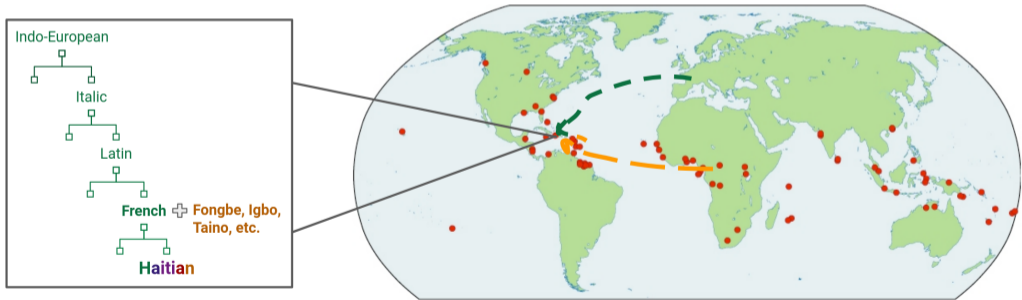
3. Outlook & Conclusion

- TrustLLM
- Example: Factual correctness
- Conclusion

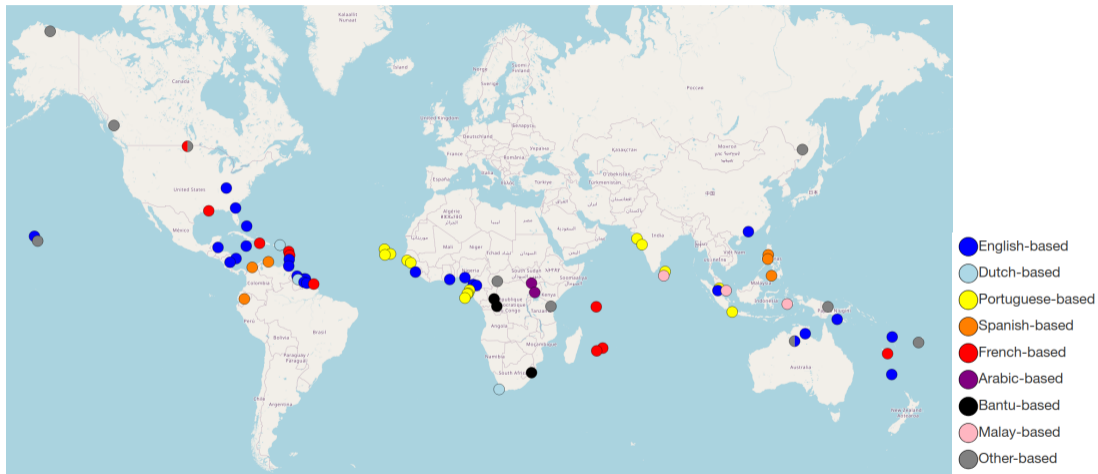
What are Creoles and why work on them?



- **Creole languages** can be found all around the world!
- Arose from **linguistic contact** between diverse languages
 - Often during Western imperialism, Atlantic slave trade
 - Inherited *stigma* related to these atrocities



Creole languages around the world



Source: APiCS

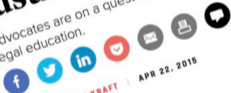
Stigmatization of creoles

- **Individual:** Some speakers don't recognize their own language
 - “*Broken English*”
- **Societal:** Possible discrimination for not speaking a “*prestige language*”
- **Institutional:** Varying amount of government recognition
 - *Lack of education* in one's native language
 - Official status in Haiti 🇧🇩, Papua New Guinea 🇵🇬, Seychelles 🇸🇪, Central African Republic 🇨🇦, Vanuatu 🇻🇺



Establishing the Rule of Law in a Country Where Justice Hardly Exists

Advocates are on a quest to improve the quality of life in Haiti through legal education.



JESSICA CAREW KRAFT | APR 22, 2015

En-jistis an franse



French is the primary language of the courts in Haiti, but as much as 97 percent of the population are *fully fluent in Creole only*, so most defendants—if they can even afford to hire a lawyer—*cannot fully grasp what goes on during the court proceedings*.

Slide credit: MIT Ayiti

Why work on Creoles?

1 Sociopolitical reason

- Reducing stigma by legitimising Creoles (Lent et al., 2022a)
- Decolonising NLP (Bird, 2020)

2 Diplomatic NLP reason

- Equal language technology for all
- Some Creoles serve as effective contact languages (Bird, 2022)

3 (Exciting!) Scientific reason

- Can lead to better NLP for all resource-poor situations
- Unique genealogical situation → lends itself for study of transfer learning!

The case for transfer learning

- Cross-lingual **vocabulary** (Lent et al., 2021)

	Tamil	Mandarin(我们)	Cantonese(拍拖)	English	Malay	Eng	Malay	Hokkien/ Hakka(店)	X
Singlish	Dey	wǒ men	paktor	always	makan	at	kopitiam	one	
	Hey	, we	date	always	eat	at	coffee shop	<INTJ>	

Standard English: "Hey, when we date we always eat at the coffee shop"

- Cross-lingual **grammar**


Spanish	Los hombres	vieron	un árbol de plátano	S	V	O
Tagalog	Nakita	ng mga lalaki	ang isang puno ng saging	V	S	O
Chavacano	Ya-mirá	el mga ómbre	un póno de ságing	V	S	O

"The men saw a banana tree."

So where can we get Creole language text?

Web crawling? Wikipedia?

What about web-crawled data?

-  **OSCAR**: web-crawled data for 150+ languages

46	gu	Gujarati	425,552	417,001,705	5.6 GB
47	ht	Haitian Creole	2	20,671	93.1 kB
48	he	Hebrew	3,997,888	1,697,158,891	18.0 GB



- Massively multilingual datasets often have **quality issues**
 - Poorest quality observed for minority languages “closely related to higher-resource languages” (Kreutzer et al., 2022)

What about Wikipedia?

Hadley, Massachusetts

Depi Wikipediya, ansiklopedi lib

Hadley, Massachusetts se yon vil [Etazini](#). Li sitye nan leta [Massachusetts](#). Chèf-lye li se York.

Albertville, Alabama

Depi Wikipediya, ansiklopedi lib

Albertville, Alabama se yon vil [Etazini](#). Li sitye nan leta [Alabama](#). Chèf-lye li se Marshall .

Grover, Kolorado

Depi Wikipediya, ansiklopedi lib

Grover, Kolorado se yon vil [Etazini](#). Li sitye nan leta [Kolorado](#). Chèf-lye li se ? .

Ball, Lwizyana

Depi Wikipediya, ansiklopedi lib

Ball, Lwizyana se yon vil [Etazini](#). Li sitye nan leta [Lwizyana](#). Chèf-lye li se ?.

West Brookfield, Massachusetts

Depi Wikipediya, ansiklopedi lib

West Brookfield, Massachusetts se yon vil [Etazini](#). Li sitye nan leta [Massachusetts](#). Chèf-lye li se York.

Quality issues of Wikipedia data

- Many articles follow simple **templates**.
 - Even longer articles often consist of lists or foreign-language names and titles.
- Some Creole Wikipedias were found to be **not at all representative** of the language by native speakers (Lent et al., 2024)
 - Chavacano: “rather an approximation of Spanish”
 - Jamaican: “spelling and grammar greatly divergent” from real-world use

Arrête-moi si tu peux

Depi Wikipedya, ansiklopedi lib

Arrête-moi si tu peux (nan angle : *Catch Me If You Can*) se yon **film** ameriken reyalize pa **Steven Spielberg** soti an 2002. Fim sa a enspire pa lavi a **Frank Abagnale Jr.**

Kontni [kache]

- 1 Ekip teknik
- 2 Aktè
- 3 Referans
- 4 Lyen deyò

Ekip teknik [modifye | modifye kòd]

Aktè [modifye | modifye kòd]

- **Leonardo DiCaprio** : **Frank Abagnale Jr**
- **Tom Hanks** : Carl Hanratty, ajan **FBI**, anketè
- **Christopher Walken** : Frank Abagnale, Sr., papa Frank
- **Nathalie Baye** : Paula Abagnale, manman Frank
- **Amy Adams** : Brenda Strong
- **Martin Sheen** : Roger Strong, papa Brenda
- **James Brolin** : Jack Barnes, prezidan klib
- **Brian Howe** : ajan **FBI** Earl Amdursky
- **Frank John Hughes** : ajan **FBI** Tom Fox
- **Steve Eastin** : Paul Morgan, le responsable de la Pan Am
- **Chris Ellis** : l'agent du **FBI** Witkins
- **John Finn** : Marsh, l'assistant du directeur du **FBI**

CreoleVal: Multilingual Multitask Benchmarks for Creoles

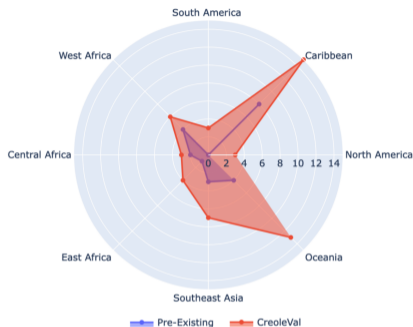


- **CreoleVal** compiles a benchmark dataset covering **up to 28 Creole languages**.

- Combines pre-existing datasets with new & *high-quality* data
- Collaboration with MIT-Ayiti

New tasks

- 1 Machine reading comprehension (n=2)
- 2 Relation classification (n=4)
- 3 Machine translation
 - MIT-Ayiti (n=1)
 - Bible (n=28)



Paper: [Lent et al. \(2024\)](#)
Data: github.com/hclent/CreoleVal

CreoleVal: Existing and new datasets

Universal Dependencies (POS)	Singlish, Naijá
Named Entity Recognition	Bislama, Chavacano, Haitian, Pijin, Papiamentu, Sango, Tok Pisin, Naijá
Sentiment Analysis	Naijá
Natural Language Inference	Jamaican
Sentence Matching	Chavacano, Guadeloupean, Haitian, Jamaican, Papiamentu, Sango, Tok Pisin
Machine Translation	Morisien
Machine Comprehension	Haitian, Morisien
Relation Classification	Bislama, Chavacano, Jamaican, Tok Pisin
Machine Translation	Haitian (<i>MIT-Ayiti</i>), + 28 more (<i>Bible</i>)

Dataset 1

Machine Comprehension

Machine Comprehension: MCTest

“James the Turtle was always getting in trouble. [...] One day, James thought he would go into town and see what kind of trouble he could get into. He went to the grocery store and pulled all the **pudding** off the shelves and ate two jars. Then he walked to the fast food restaurant and ordered 15 bags of fries. He didn't pay, and instead headed home.

What did James pull off of the shelves in the grocery store?

- A) **pudding**
- B) fries
- C) food
- D) town



Adapted from MCTest

Translating MCTest to Creole languages

💰 We have some money. Now what?

- 1 Pick a small dataset in English, e.g. MCTest160 dev set (30 stories, 120 questions)
- 2 Pay translators to translate them into Creole languages.

🚧 **Problem:** Translators for Creoles are **difficult to find!**

- Found professional translators for Haitian Creole & Mauritian Creole (Morisien)
- *We could have hired more!* 🙄



Greta ran to the corner with her older brother Tony . He had money for the ice cream truck in his pocket and she was very happy.



Greta te kouri ale nan kafou a avèk gran frè I, Tony . Li te gen lajan pou achte nan kamyon krèm lan nan pòch li epi Greta te kontan anpil.



Agat te kouri ale nan kafou a avèk gran frè I, Toni . Li te gen kòb nan pòch li pou te achte nan men machann fresko lan nan pòch li epi Agat te kontan anpil.



❖ **Problem:** Literal translations can result in **culturally irrelevant** content.

→ We created both *direct* and *localized* translations!

Dataset 2

Relation Classification

Relation Classification

date of birth

Mark Twain was born in 1835.

place of birth

Elvis Presley was born in Memphis, Tennessee.

may treat

Ribavirin remains essential to Hepatitis C treatment.

Examples from Gao et al. (2019)

[CITY] se yon vil [COUNTRY].

Hadley, Massachusetts

Depi Wikipedya, ansiklopedi lib

Hadley, Massachusetts se yon vil [Etazini](#). Li sitye nan leta [Massachusetts](#). Chèf-lye li se York.

Albertville, Alabama

Depi Wikipedya, ansiklopedi lib

Albertville, Alabama se yon vil [Etazini](#). Li sitye nan leta [Alabama](#). Chèf-lye li se Marshall .

Grover, Kolorado

Depi Wikipedya, ansiklopedi lib

Grover, Kolorado se yon vil [Etazini](#). Li sitye nan leta [Kolorado](#). Chèf-lye li se ? .

Ball, Lwizyana

Depi Wikipedya, ansiklopedi lib

Ball, Lwizyana se yon vil [Etazini](#). Li sitye nan leta [Lwizyana](#). Chèf-lye li se ?.

West Brookfield, Massachusetts

Depi Wikipedya, ansiklopedi lib

West Brookfield, Massachusetts se yon vil [Etazini](#). Li sitye nan leta [Massachusetts](#). Chèf-lye li se York.

Building RC datasets for Creoles

- The **latent templates** give us candidate sentences expressing relations.

Berlin a di capital fi Germany.

Athens a di capital fi Greece.

Sofia ah di capital fi Bulgaria.

Bogota a di capital fi Colombia.

Daka a di capital a di biggest city inna Bangladesh.

Q64 — P1376 → Q183



- Relation annotation** is relatively easy on this data!
 - Quality assurance: Creole native speakers verify the sentences & annotations.

Things we did *not* do in CreoleVal...

What about a Creole language model?

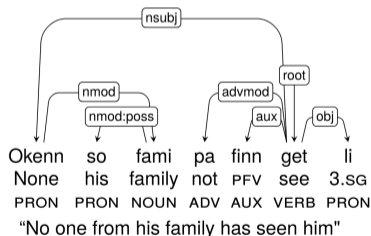
- We played around with the idea of training a **CreoleLM**.
 - 🤔 Could fine-tune an existing, multilingual model (e.g. mBERT)
 - 🤔 Could train a small *language adapter* (e.g. for BLOOM)
- Ultimately, we decided we still had **too little data** for this. 🤔
 - Ideally, we'd want *longer, coherent* texts.
 - Most of our long-form texts are Bible translations
→ not ideal for a general-purpose LM...

What about treebanks?

- Treebanks could be useful for **spell & grammar checkers**.
 - Community goal: Help foster **a culture of writing!**
- They are also **expensive to produce**.
 - And you need to find the experts...

Recent example

- **Ramsurrun et al. (2024)** construct a 161-sentence treebank for Mauritian Creole.



Contributors to CreoleVal



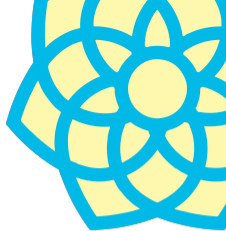
An outlook

(on TrustLLM)

& a conclusion



Trustworthy LLMs for Europe



- TrustLLM is an ongoing, EU-funded project to **develop LLMs** for the **Germanic languages**.
 - Collaboration between 11 partners (all in Germanic-speaking countries)
 - Lowest-resource language “officially” in the project: **Icelandic!**
 - TrustLLM aims to develop models that are open, sustainable, and **trustworthy**.
 - e.g. factual correctness, consistency, non-discrimination
- We need good **evaluation datasets** to assess our trustworthiness criteria!


Example: Factual correctness (MMLU)

“ *Which element in tobacco smoke is responsible for cancers?*

- A) *Nicotine*
- B) *Tar*
- C) *Carbon monoxide*
- D) *Smoke particles*

”

Idea

- 1 Machine-translate  **MMLU** into several Germanic languages.
- 2 Use native speakers within our project to check & correct the data.

Pitfalls of translation, again

- ⚡ **Problem:** Many questions in MMLU are highly **culturally specific** to the US.
 - e.g. categories “US foreign policy”, “professional law”, “moral disputes”, ...

“

The attorney for a plaintiff in an action filed in a federal district court served the defendant with the summons, the complaint, and 25 interrogatories asking questions about the defendant's contentions in the case. The interrogatories stated that they were to be answered within 30 days after service. The defendant is likely to succeed in obtaining a protective order on which of the following grounds?

”



Global MMLU :
Understanding and
Addressing Cultural and
Linguistic Biases in
Multilingual Evaluation



“ We found out that 28% of MMLU requires cultural context to be answered correctly. An outstanding 85% are tagged as specific to Western culture or western regions. **Progress on MMLU requires excelling at western culture.** ”

— Sara Hooker on Bluesky, yesterday!

and better

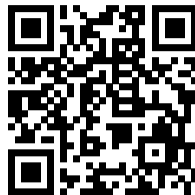
We always need more data!

- Training & fine-tuning models
- Benchmark evaluations
- Analysing model capabilities
- More variety of tasks
- More language coverage
- Culturally appropriate data

Thank you!

🏠 marcel.bollmann.me
✉ marcel.bollmann@liu.se
🦋 [@bollmann.me](https://twitter.com/bollmann.me)

CreoleVal 
github.com/hclent/CreoleVal



 TrustLLM
trustllm.eu