Automatic Speech Recognition and Query By Example for Creole Languages Documentation

Cécile Macaire[1], Didier Schwab[1], Benjamin Lecouteux[1], Emmanuel Schang[2]

[1]Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, 38000 Grenoble, France
[2]LLL, UMR 7270, Univ. Orléans & CNRS

[1]*first.last@univ-grenoble-alpes.fr*, [2]*emmanuel.schang@univ-orleans.fr*

ACL 2022
22ND – 27TH MAY | 60TH MEETING | DUBLIN

What is it about ?

Currently, linguists gather several hours of speech.
The task of documenting Creole languages, such as:

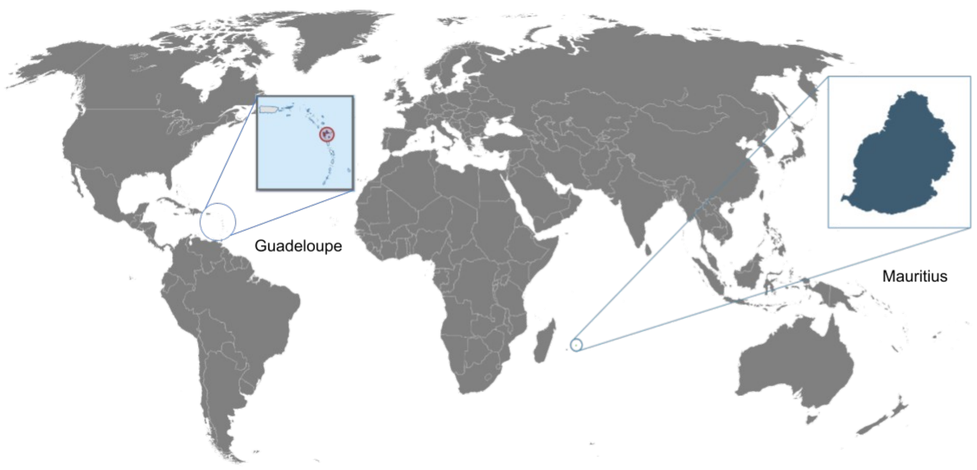
- **transcribing the speech**,
- **search for a specific word** in a set of audio recordings,

is time-consuming for field linguists.

How to simplify the documentation of Creole languages and help the work of linguists?

Transcribing Creole languages

Creole languages remains, for many of them, **under-resourced languages**.
This work focuses on two Creole languages: **Gwadloupéyen** & **Morisien**.



Few challenges:

- Largely under-equipped languages which suffer from a low social status.
- Mostly spoken **used in context of a dominant language**:
  French for Gwadloupéyen [1, 2], French and English for Morisien (most of the words are easily identifiable by a French speaker) [3, 4].
- **Unstable written form** (words transcribed in two different forms, transcription in proper Creole form).
- **Code-switching** [1, 5].

From the Gwadloupéyen:

(2) modes de cuisson qui adaptés osi
fr fr fr fr fr cr
methods of cooking that adapt too
“cooking methods which are adapted too”

→ **Hesitation** between a transcription in French or in Creole.

What we propose

To efficiently correct the transcriptions errors and allow the linguist to search for a specific word in a corpus independently of its transcription, we:

(1) Use about one hour of annotated data to **design an automatic speech recognition (ASR) system** for each language based on self-supervised learning (SSL).
Self-supervised learning (SSL) is the task of learning powerful representations from huge unlabeled data (called pretraining) to recognize and understand patterns from a less common problem (called fine-tuning).
→ **effective performances on downstream tasks** for ASR in low-resource contexts [6, 7] .

(2) Design a **Query-by-Example (QbE) system** to allow the linguist to verify or correct the transcription.
Query by Example (QbE) detects specific words in speech recordings thanks to the use of speech recognition approaches.
Two steps:

- using self-supervised learning models to predict the word in a speech segment,
- search the prediction in a set of transcriptions.

Methodology

Datasets / Gwadloupéyen - gcf, 80 min and 5 speakers.
Morisien - mfe, 60 min and 2 speakers.
→ Spontaneous speech

Pre-processing for fine-tuning / Audio recordings split into small segments.
Remove specific characters from the textual data.

Implementation details / Fine-tuning performed using two pretrained models:

- *XLSR-53* [8], a multilingual pretraining of Wav2Vec2.0 model on 53 languages with more than 56k hours of unlabeled speech data.
- *LeBenchmark* [9] , a French-based Wav2Vec2.0 model.

Decoding with a CTC beam search decoder + LM.

QbE / Extraction of speech segments corresponding to a Gwadloupéyen word.
→ based on the Smith-Waterman algorithm (output an optimal alignment between two given sequences by looking at matching areas) [10].

Results

Model	Training size (in min)	Pretrained model	LM	dev		test	
				WER (%)	CER (%)	WER (%)	CER (%)
gcf_xlsr	68	facebook/wav2vec2-large-xlsr-53	-	47.58	22.60	40.68	17.81
			3-gram	-	-	37.91	18.59
gcf	68	LeBenchmark/wav2vec2-FR-7K-large	-	39.50	17.89	35.96	15.86
			3-gram	-	-	**34.74**	**16.96**
mfe_xlsr	52	facebook/wav2vec2-large-xlsr-53	-	48.08	21.56	44.66	20.06
			3-gram	-	-	41.60	20.12
mfe	52	LeBenchmark/wav2vec2-FR-7K-large	-	41.44	18.23	**36.19**	**16.70**
			3-gram	-	-	38.83	18.03

Table 1: Word Error Rate (WER) and Character Error Rate (CER) on different creole languages when finetuning the Wav2Vec2.0 model with multilingual (*XLSR-53*) and monolingual (*LeBenchmark/wav2vec2-FR-7Klarge*) models. The WER and the CER are given with and without a 3-gram LM on the test sets.

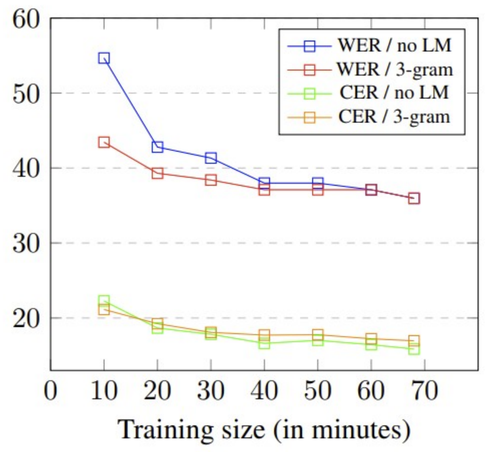


Figure 1: WER and CER (%) with respect to different training sizes (in minutes) when fine-tuning *LeBenchmark* pretrained model on the Gwadloupéyen corpus. The WER and the CER are given on the test sets with (in red and orange) or without a 3-gram LM (in blue and green)

Fine-tuned model	Precision (%)	Recall (%)	F-measure (%)
gcf_10	83.33	74.36	78.59
gcf_20	83.33	76.28	79.65
gcf_30	72.50	79.49	75.83
gcf_40	75.00	74.36	74.68
gcf_50	66.67	66.67	66.67
gcf_60	**84.52**	**84.94**	**84.73**

Table 3: Precision, Recall and F-measure computed on the Qbe results of 13 Gwadloupéyen audio segments when using the fine-tuned gcf models trained with 10 (gcf_10) to 60 minutes (gcf_60) of training data to predict the utterance. Audio segments contain single words (e.g. ‘dépi’, ‘fè’) and multiple words (e.g. ‘an pa sav’, ‘nou ka rivé’)

Prediction from the Gwadloupéyen ASR system:

```
Ref:   sé timoun    pa ni pon rèspè
Hyp:   sé timoun la pa ni pon respè
```

Discussion

On the Automatic transcription /

- **Accurate enough to allow for a fast manual correction.**
- Better results with the French pretrained model which sheds new light on the question of the link between a Creole language and the so-called ‘lexifier’ language.

On the Qbe /

- **Complement ASR, provides an easy way to scan the corpus for relevant examples.**

Acknowledgments

The CREAM project (Documentation des Langues CREoles Assistée par la Machine) is funded by the ANR (Agence Nationale de la Recherche, CS-38, 2020-2024). We would like to thank Dr. T. Veenstra (ZAS, Berlin) and Dr. F. Henri (University at Buffalo) for sharing with us their data on Morisien.

References

[1] Hazaël-Massieux, G. (1978). Approche socio-linguistique de la situation de diglossie français-créole en Guadeloupe. Langue française, (37), 106-118.
[2] Managan, J. K. (2004). Language choice, linguistic ideologies and social identity in Guadeloupe. New York University.
[3] Boswell, R. (2006). Le malaise créole: ethnic identity in Mauritius (Vol. 26). Berghahn Books.
[4] Rajah-Carrim, A. (2005). Language use and attitudes in Mauritius on the basis of the 2000 population census. Journal of multilingual and multicultural development, 26(4), 317-332.
[5] Auckle, T. (2015). Code switching, language mixing and fused lects: language alternation phenomena in multilingual Mauritius (Doctoral dissertation).
[6] Baevski, A., Auli, M., & Mohamed, A. (2019). Effectiveness of self-supervised pre-training for speech recognition. arXiv preprint arXiv:1911.03912.
[7] Kawakami, K., et al. (2020). Learning robust and multilingual speech representations. In Findings of the Association for Computational Linguistics: EMNLP 2020, pages 1182–1192, Online. Association for Computational Linguistics.
[8] Conneau, A., et al.. (2021). Unsupervised Cross-Lingual Representation Learning for Speech Recognition. In Proc. Interspeech 2021, pages 2426–2430.
[9] Evain et al. (2021). LeBenchmark: A Reproducible Framework for Assessing Self-Supervised Representation Learning from Speech. In Proc. Interspeech 2021, pages 1439–1443.
[10] Lecouteux, B., et al.. (2012). Integrating imperfect transcripts into speech recognition systems for building high-quality corpora. Computer Speech & Language, 26(2), 67-89.