Comparing Yoruba and French Languages in Low-Resource Domain for Automatic Speech Recognition

Sewade Olaolu Ogun, Tatiana Moteu

Contents

1 Introduction

Automatic speech recognition (ASR) for high-resource languages is commonplace, but low-resource ASR is gaining interest in the community as humans (infants) require few examples to learn speech and language. Low-resource ASR requires learning with few examples which is an interesting area of research. This project aims to explore transfer learning for low-resource ASR by utilizing a pretrained Contrastive Predictive Coding(CPC) model and fine-tuning it on Yoruba and French Language ASRs given only 1 hour of data for each language.

2 Data

2.1 Data Preparation

Recording was done using the Lig-Aikuma Android app. It is an easy-to-use app with a good interface for recording and elicitation. In the elicitation mode, a text is displayed on screen while the speaker reads out the text carefully. Dataset for the experiments are hosted at [1] and [2] for both yoruba and french datasets.

2.2 Data Preprocessing

Raw speech dataset is split into;

- 1 hour of speech data (split into train and val (80%-20% split)
- 1 hour of speech as test set

The audio files are then preprocessed with their respective text files. Text files are linked to the audio files and converted to character strings as required by the CPC model.

3 Summary of Results

The CPC model was initially trained without starting with pretrained weights to see how it will perform with few examples. Character Error Rates (CER) are measured for all the experiments as well training/validation losses. The pretrained CPC model is gotten from the CPC audio library from Facebook research [3].

Table 1 summarizes results of experiments comparing both languages when trained from scratch and when fine-tuned using a pretrained model. After 80 epochs, the CER was still very high without pretraining. By fine-tuning a pretrained model for the same experiment, we were able to reduce the word-error-rate by a factor of 2. For the yoruba experiments, the model indicates it can get better but will require more training time to converge to a reasonable result.

For all experiments, we use the same hyper-parameter configuration of

- learning rate: 2e-4
- number of epochs: 80 epochs,
- Early stopping,
- Beam search - beam width: 20, cutoff top n: 20

Training loss plateau for up to 20-30 epochs for most experiments before decreasing considerably for the remaining epochs. The ReduceOnPlateau learning rate scheduler helped in ensuring training continues.

Furthermore, we experimented with 4-gram and 5-gram language models for the yoruba language CPC model. The language models were trained using the kenlm library following the procedure in [4]. Table 2 shows the results of all experiments performed for yoruba. The result show that language models can serve as a good prior to the beam search. 5-gram language models performed better than 4-gram language models on the test set while further performing better than models not using any language model. Table 3 shows the results of experiments performed on the french dataset. In comparison with the yoruba models, the WER on the french dataset was not as low as that of yoruba language on the same task. This might indicate that yoruba language is closer to english in language proximity than French.

Table 1: Comparison of Results for Yoruba and French Models

	Yoruba	**French**
Model	**WER (Test)**	**WER (Test)**
CPC trained from scratch	0.6258	0.7177
CPC Finetuned without LM	0.3476	0.4792

Table 2: Training Results for Yoruba Language Experiments

	Train	**Val**		**Test**
Model	**Loss**	**Loss**	**CER**	**CER**
CPC trained from scratch	1.8845	2.0444	0.6053	0.6258
CPC Finetuned without LM	0.4719	1.1641	0.3040	0.3476
CPC with 4-Gram LM	0.4885	1.2059	0.3072	0.3395
CPC with 5-Gram LM	0.3208	1.2234	0.3040	0.3270

Table 3: Training Results for French Language Experiments

	Train	**Val**		**Test**
Model	**Loss**	**Loss**	**CER**	**CER**
CPC trained from scratch	2.6973	2.6705	0.7108	0.7177
CPC Finetuned without LM	0.7225	1.933	0.4792	0.4792

4 Future Directions

Some other interesting ideas which we could not explore at the moment include;

- Increase the training data in steps and observing the WER due to data size
- Multilingual finetuning using both french and yoruba dataset
- Use RNN language model or Transformer-based language model as language prior.
- Multilingual pretraining (using CPC).

5 Challenges

The major challenge faced was in creating and preparing the datasets for this experiment. Getting the data in the format required for dataloaders require that scripts had to be written to place the data in the right format.

6 Conclusion

In this project, we built an Automatic speech recognition system (character level) for both yoruba language and french language with only 1 hour of labeled data. Then, we compared the results of the models to show thier capacity in low resource settings. The results show that pre-trained models can converge faster than training from scratch as the optimization is more well behaved, and should be explored for tranfer learning to other languages with low resources.

References

[1] Sewade, O. 2020, Yoruba Speech Dataset for Low-Resource Speech Tasks, https://github.com/ogunlao/yoruba_speech_project

[2] Tatiana, M. 2020, French Speech Rec, https://github.com/TatianaMoteuN/french_speech_rec

[3] Morgane Rivière & Armand Joulin & & Pierre-Emmanuel Mazaré & Emmanuel Dupoux 2020, Unsupervised pretraining transfers well across languages, arXiv 2002.02848

[4] Kenneth Heafield, KenLM Language Model Toolkit, https://kheafield.com/code/kenlm/, Retrieved July 2020

A CPC Train and Validation Loss for Yoruba

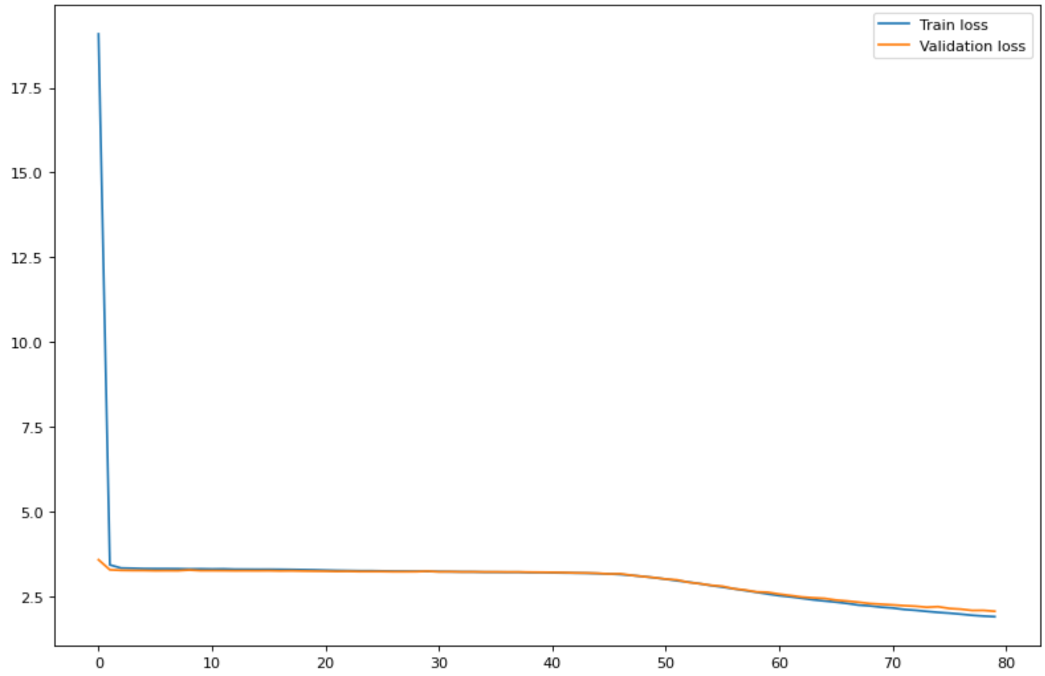

Figure 1: Train and Validation Loss for CPC Trained from Scratch and No Language Model

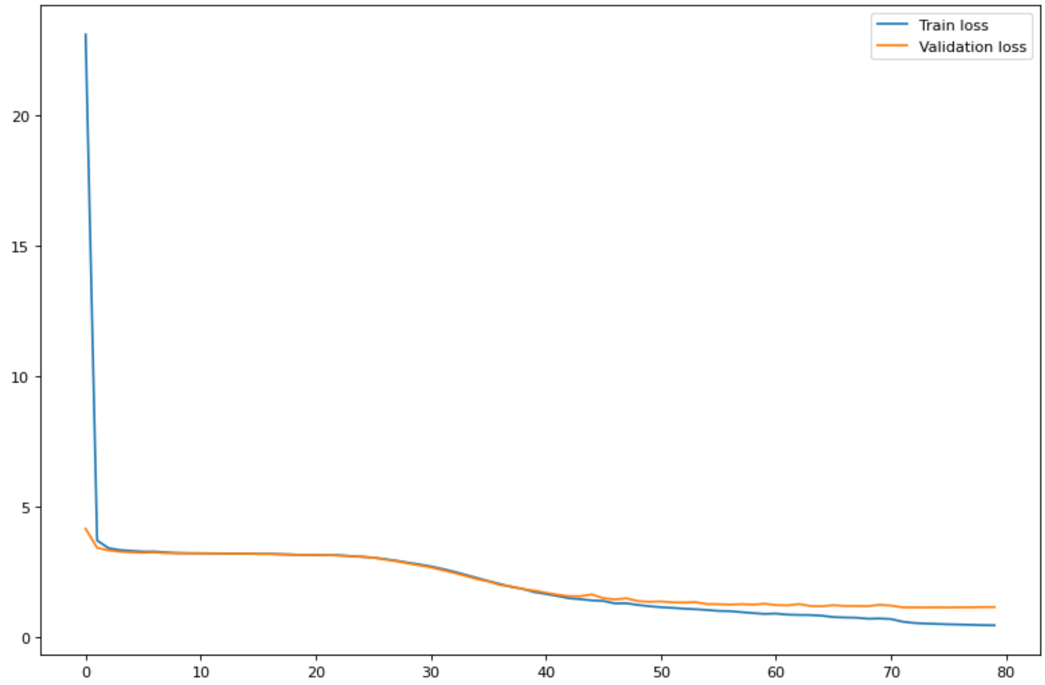

Figure 2: Train and Validation Loss for CPC with Fine Tuning and No Language Model

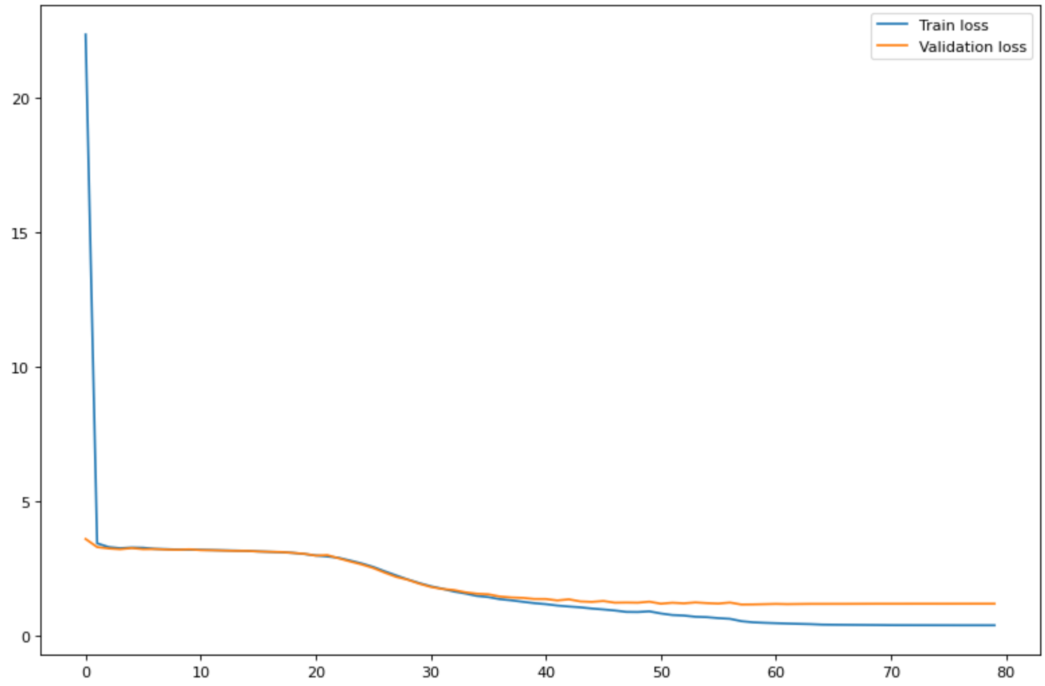

Figure 3: Train and Validation Loss for CPC with Finetuning and 4-Gram Language Model

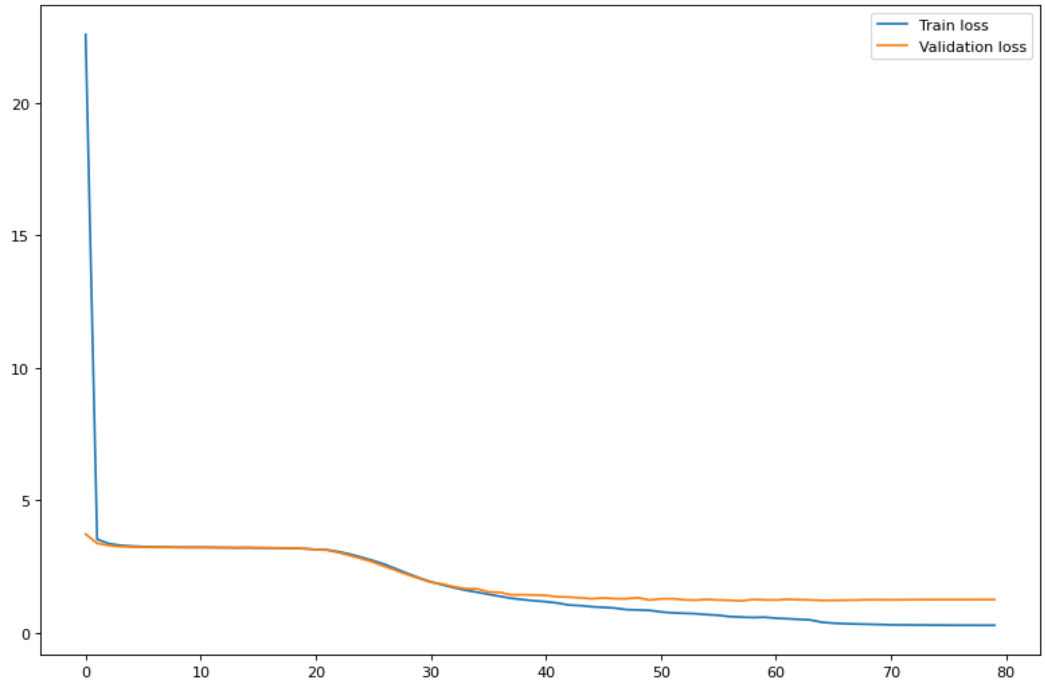

Figure 4: Train and Validation Loss for CPC with Finetuning and 5-Gram Language Model

B CPC Train and Validation Loss for French

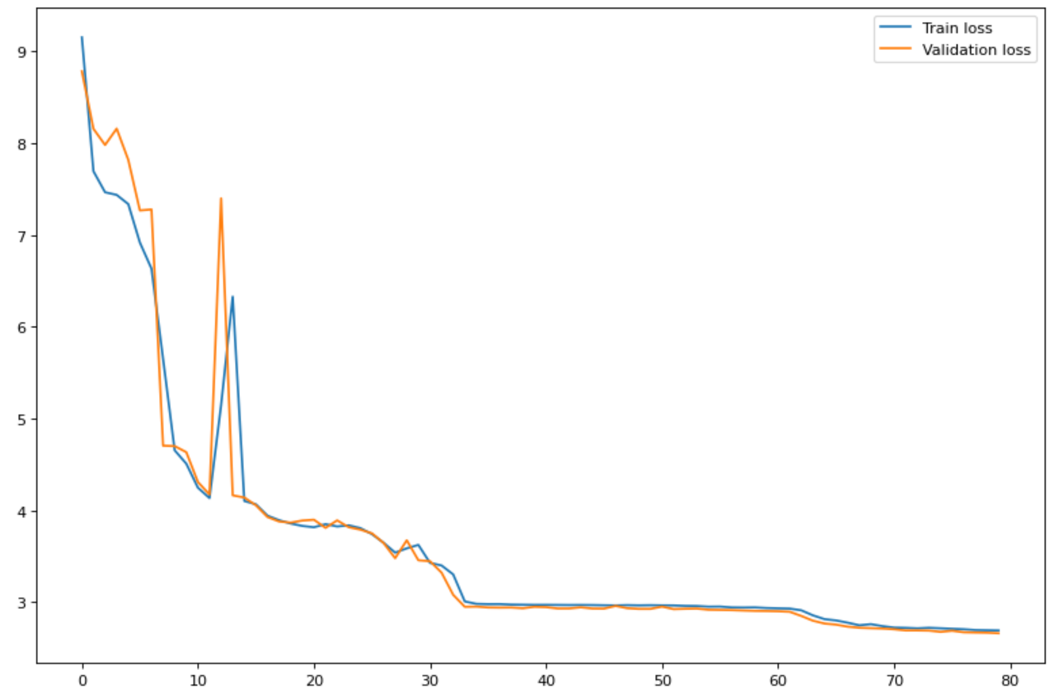

Figure 5: Train and Validation Loss for CPC From Scratch - No Language Model

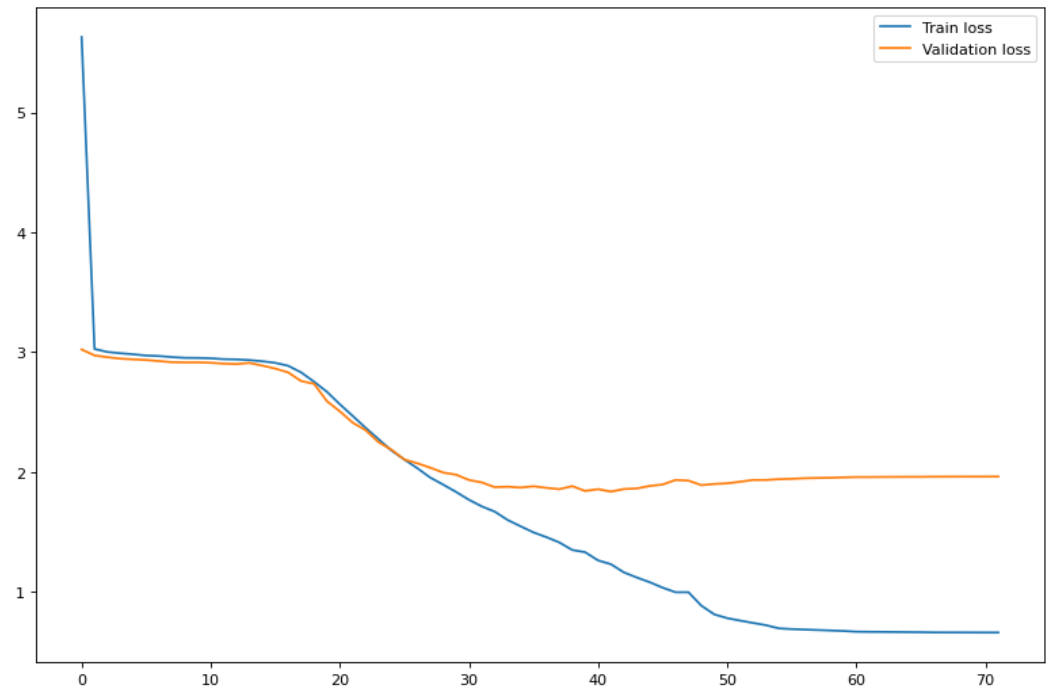

Figure 6: Train and Validation Loss for CPC with Finetuning - No Language Model

C Character Error Rate (CER) Plots for Yoruba

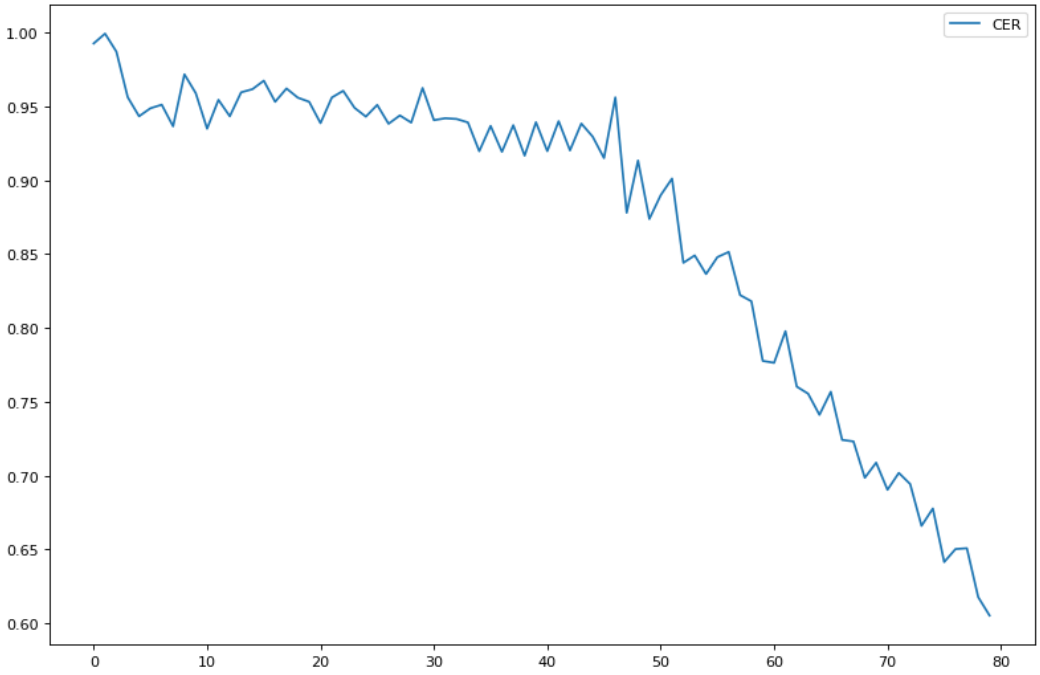

Figure 7: CER Plot for CPC Trained From Scratch

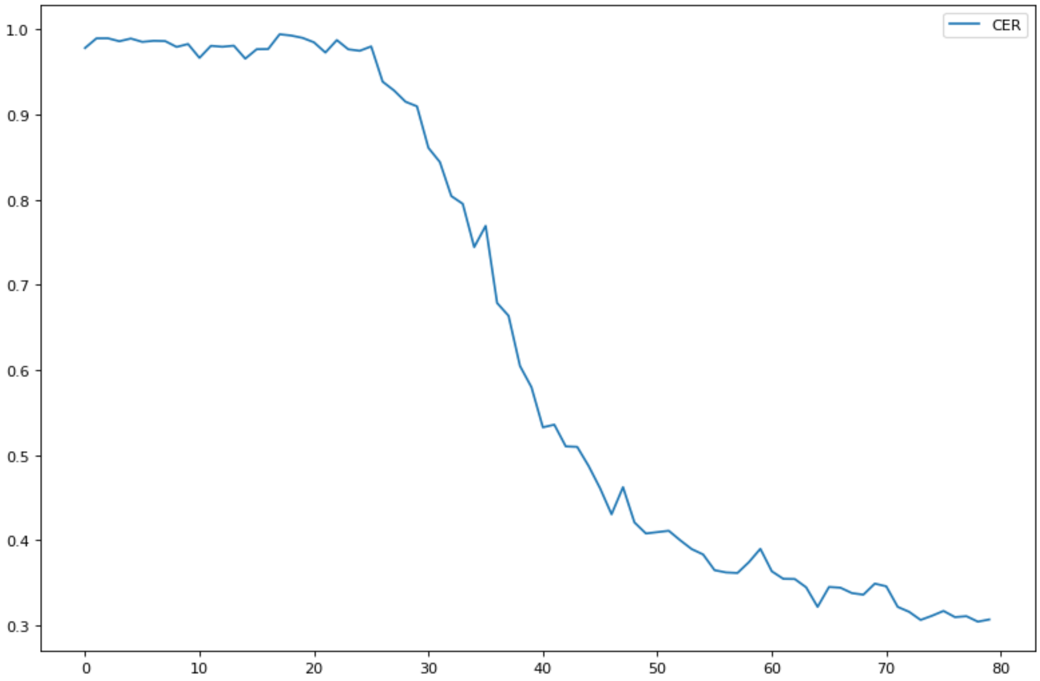

Figure 8: CER Plot for CPC with No Language Model

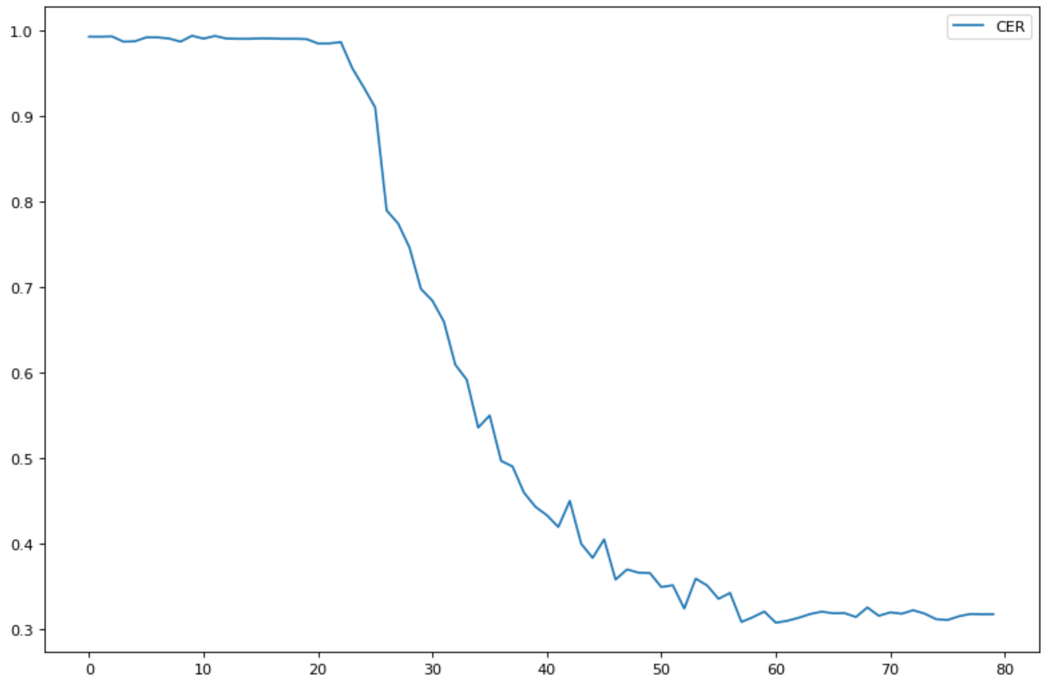

Figure 9: CER Plot for CPC with 4-Gram Language Model

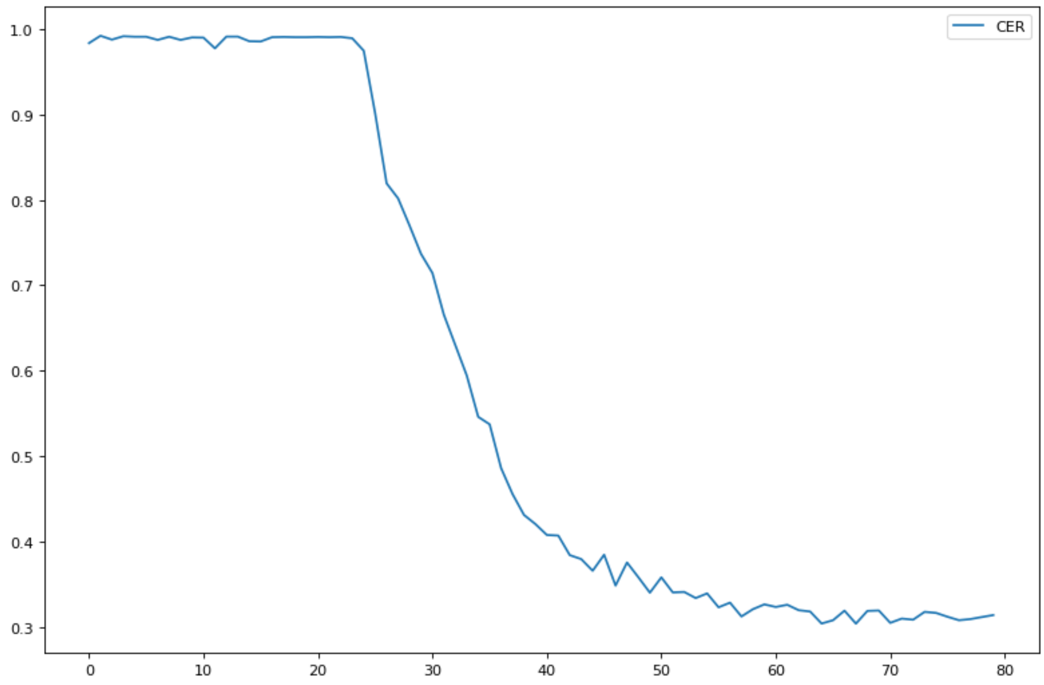

Figure 10: CER Plot for CPC with 5-Gram Language Model

D Character Error Rate (CER) Plots for French

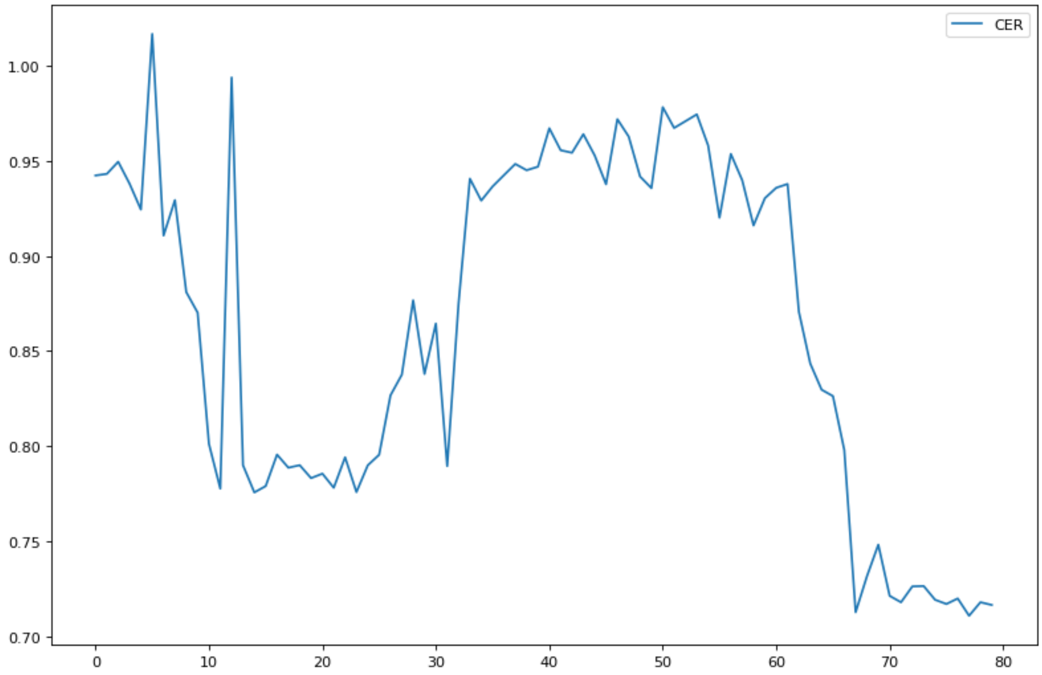

Figure 11: CER Plot - CPC from Scratch with No Language Model

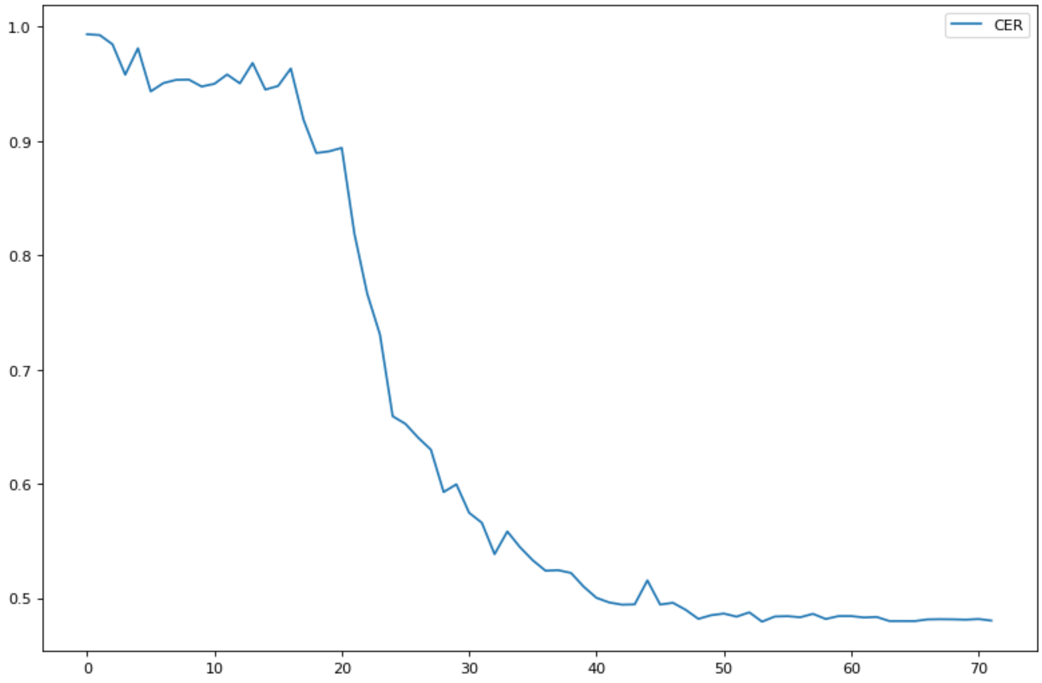

Figure 12: CER Plot - CPC with No Language Model