Ancestor-to-Creole Transfer is Not a Walk in the Park

Lent, Heather; Bugliarello, Emanuele; Søgaard, Anders

Published in:
Proceedings of the Third Workshop on Insights from Negative Results in NLP

DOI:
10.18653/v1/2022.insights-1.9

Publication date:
2022

Document version
Publisher’s PDF, also known as Version of record

Document license:
CC BY

Citation for published version (APA):
Ancestor-to-Creole Transfer is Not a Walk in the Park

Heather Lent  Emanuele Bugliarello  Anders Søgaard
University of Copenhagen, Denmark
{hcl, emanuele, soegaard}@di.ku.dk

Abstract

We aim to learn language models for Creole languages for which large volumes of data are not readily available, and therefore explore the potential transfer from ancestor languages (the ‘Ancestry Transfer Hypothesis’). We find that standard transfer methods do not facilitate ancestry transfer. Surprisingly, different from other non-Creole languages, a very distinct two-phase pattern emerges for Creoles: As our training losses plateau, and language models begin to overfit on their source languages, perplexity on the Creoles drop. We explore if this compression phase can lead to practically useful language models (the ‘Ancestry Bottleneck Hypothesis’), but also falsify this. Moreover, we show that Creoles even exhibit this two-phase pattern even when training on random, unrelated languages. Thus Creoles seem to be typological outliers and we speculate whether there is a link between the two observations.

1 Introduction

Creole languages refer to vernacular languages, many of which developed in colonial plantation settlements in the 17th and 18th centuries. Creoles most often emerged as a result of contact between social groups that spoke mutually unintelligible languages, i.e., from the interactions of speakers of nonstandard varieties of European languages and speakers of non-European languages (Lent et al., 2021). Some argue these languages have an exceptional status among the world’s languages (McWhorter, 1998), while others counter that Creoles are not unique, and evolve in the typical manner as other languages (Aboh and DeGraff, 2016).

In this paper, we will present experiments in evaluating language models trained on non-Creole languages for Creoles, as well as in various control settings. We first explore the following hypothesis:

R1: Language models trained on ancestor languages should transfer well to Creole languages.

We call R1 the ‘Ancestry Transfer Hypothesis.’ Our experiments, however, suggest that R1 is not easily validated. We note, though, that ancestor-to-Creole training exhibits divergent behavior when training for long, leading to the following hypothesis:

R2: Language models trained on ancestor languages can, after a compression phase, transfer well to Creole languages.

We call R2 the ‘Ancestry Bottleneck Hypothesis.’ While compression benefits transfer, performance never seems to reach useful levels. Furthermore, similar effects are observed with Creoles when training on non-ancestor languages. Our findings here are not relevant to applied NLP, but they shed light on cross-lingual training dynamics (Singh et al., 2019; Deshpande et al., 2021), and we believe they have potential implications for the linguistic study of Creoles (DeGraff, 2005b), as well as for information bottleneck theory (Tishby et al., 1999).

Our contributions We conduct a large set of experiments on cross-lingual zero-shot applications of language models to Creoles, primarily to test whether ancestor languages provide useful training data for Creoles (the ‘Ancestry Transfer Hypothesis;’ R1). Our results are a mix of negative and positive results: First Negative Result: Ordinary transfer methods do not enable ancestor-to-Creole transfer. First Positive Result: Regardless of the
source languages, when training for long periods of time, a compression phase takes places for Creoles: as the models overfit their training data, perplexity on Creoles begin to decrease. This pattern is unique to Creoles as it does not emerge for target non-Creole languages. **Second Negative Result:** The compression phase does not lead to better representations for downstream tasks in the target Creoles.

## 2 Background

**Cross-lingual training dynamics** Several multilingual language models have been presented and evaluated in recent years. Since Singh et al. (2019) showed that mBERT (Devlin et al., 2019) generalizes well across related languages, but compartmentalizes language families, several researchers have explored the training dynamics of training multilingual language models across related or distant language sets (Lauscher et al., 2020; Keung et al., 2020; Deshpande et al., 2021). Unlike most previous work on cross-lingual training, we focus on evaluation on unseen (Creole) languages. This set-up is also explored in previous work focusing on generalization to unseen scripts (Muller et al., 2021; Pfeiffer et al., 2021). Muller et al. (2021) argue that generalization to unseen languages is possible for seen scripts, but hard or impossible for unseen scripts, but this paper identifies a third category of unseen languages with seen scripts, which exhibit non-traditional learning curves in the zero-shot pre-training regime.

**Linguistic theories of Creole** Creolists have long debated whether Creole languages have an exceptional status among the world’s languages (DeGraff, 2005a). McWhorter (1998) argue that Creoles are *simpler* than other languages, and defined by minimal usage of inflectional morphology, little or no use of tone encoding lexical or syntactic contrasts, and generally semantically transparent derivation. Others have argued that Creoles cannot be be unambiguously distinguished from non-Creoles on strictly structural, synchronic grounds (DeGraff, 2005a). On this view Creole grammars do not form a separate typological class, but exhibit many similarities with the grammars of their parent languages, e.g., the similarities in lexical case morphology between French and Haitian Creole. We do not take sides in this debate, but observe that the exceptionalist position would explain our results that zero-shot transfer to Creole languages is particularly difficult. Exceptionalism also aligns well with the heatmaps presented in §5.

**Information Bottleneck** The Information Bottleneck principle (Tishby et al., 1999) is an information-theoretic framework for extracting output-relevant representations of inputs, i.e., compressed, non-parametric and model-independent representations that are as informative as possible about the output. Compression is formalized by mutual information with input. A Lagrange multiplier controls the trade-off between these two quantities (informativity and compression). Being able to compute this trade-off assumes the joint input–output distribution is accessible. The trade-off is found by ignoring task-irrelevant factors and learning an invariant representation. The intuition behind the ‘Ancestry Bottleneck Hypothesis’ (R2) is that invariant representations are particularly useful for Creoles (see Figure 1 for an illustration).

## 3 Multilingual Training

This section sets out to evaluate the ‘Ancestry Transfer Hypothesis’ (R1). To this end, we evaluate multilingual language models – trained with a BERT architecture from scratch, but of smaller size and with less data (Dufter and Schütze, 2020) – on Creoles such as Nigerian Pidgin or Haitian Creole. We compare two scenarios: 1) a scenario in which the training languages are languages that are
Four zero-shot transfer experiments for Creole languages. The left-hand side plot shows the (zero-shot) validation curve for checkpoints on Creole data; the small plots show the learning curves for the training languages. We see an initial increase in perplexity (disproving \textit{R1}). The yellow vertical line denotes 100 epochs. We also see a subsequent decrease in perplexity.

said to be parent or ancestor languages of the Creole, such as French to Haitian, and 2) a scenario in which random, unrelated training languages were selected. To compare against Creoles, we also explore these transfer scenarios for two target non-Creoles – Spanish and Danish – training on languages closely related to them (i.e., as typically done in cross-lingual learning). Table 1 lists all the transfer scenarios that we investigated. Our experimental protocol follows Dufter and Schütze (2020), and it is described in detail below.

We aim to learn language models for Creole languages for which large volumes of data are not readily available, and therefore explore the potential transfer from ancestor languages (the ‘Ancestry Transfer Hypothesis’). We find that standard transfer methods do not facilitate ancestry transfer. Surprisingly, different from other non-Creole languages, a very distinct two-phase pattern emerges for Creoles: As our training losses plateau, and language models begin to overfit on their source languages, perplexity on the Creoles drop. We explore if this compression phase can lead to practically useful language models (the ‘Ancestry Bottleneck Hypothesis’), but also falsify this. Moreover, we show that Creoles even exhibit this two-phase pattern even when training on random, unrelated languages. Thus Creoles seem to be typological outliers and we speculate whether there is a link between the two observations.

**Experimental protocol** We train BERT-smaller models (Dufter et al., 2020), consisting of a single attention head (shown to be sufficient for achieving multilinguality by K et al. 2020). Although training smaller models means our results are not directly comparable to larger models like mBERT or XLM-R (Conneau et al., 2019), there is evidence to support that smaller transformers can work better for smaller datasets (Susanto et al., 2019), and that the typical transformer architecture would likely be overparameterized for our small data (Kaplan et al., 2020). Thus, the BERT-smaller models appear to be the most appropriate match for our very small datasets. The models are trained on a multilingual dataset, consisting of an equal parts of each source
language, taken from the Bible Corpus (Mayer and Cysouw, 2014). We chose Bible data to train our models as it facilitates a controlled setup with parallel data in many languages whilst including our low-resource Creoles and ancestors. For each experiment, we learn a custom BERT tokenizer on source and target languages, with a vocabulary size of 10,240 word pieces (Wu et al., 2016). Each model is trained for 100 epochs (see Table 2).

We also follow Dufter and Schütze (2020)’s approach of calculating the perplexity on 15% of randomly masked tokens ($w$), with probabilities ($p$), as $\exp(-1/n \sum_{k=1}^{n} \log(p_{w_k}))$. We calculate perplexity on held out development data for both source and target languages. Our code is available online.\(^\text{2}\)

### Results

In Figure 2, by 100 epochs (indicated by a yellow vertical line), we observe two different patterns for Creoles and non-Creoles. For target Creole languages, the models are able to learn the ancestor languages, but perplexity on the held out Creoles consistently climbs. On the other hand, for target non-Creoles, we observe a slight initial drop in perplexity before it starts to increase as the models overfit the source languages.

## 4 Training For Longer

It seems linguistically plausible that training for longer on ancestor languages to learn more invariant representations should better facilitate zero-shot transfer to Creole languages. This is the essence of the ‘Ancestry Bottleneck Hypothesis’ (R2), which we explore in this section.

\(^{1}\) We explored different vocabulary sizes (1,024, 2,048 and 10,240) as well as other tokenization techniques (grapheme-to-phoneme and byte-pair encodings Sennrich et al. 2016), which did not affect the overall findings discussed below.

\(^{2}\) https://github.com/hclent/ancestor-to-creole

### Creole compression

We continue training our models for 5 days, for each Creole and non-Creole target language – which typically results in 300k–500k steps of training (and thus, extremely overfit). As the models overfit to the source languages, we observe a notable drop in perplexity for Creoles, which is true regardless of the training data (ancestors versus random controls), as shown in Figure 2 and Figure 3. On the other hand, these plots show that this compression does not emerge for non-Creole target languages, as their complexity steadily increases as the models overfit their training data more and more.

### Downstream performance

Next, in order to determine if this compression present for Creoles can be beneficial, we used MACHAMP (van der Goot et al., 2021) to check the ability of our Nigerian Pidgin models to fine-tune for downstream NER (Adejani et al., 2021). We evaluate the representations learned at different stages of pre-training by fine-tuning our checkpoints corresponding to early stage (10,000 steps), maximum perplexity, and post-compression (last checkpoint). Each model is fine-tuned for 10 epochs. Figure 4 shows that, across three random seeds, post-compression checkpoints consistently perform worse than pre-compression or max-complexity checkpoints. The results negate R2, i.e., that the compression effect observed during training would be useful for Creoles.\(^3\)

### Few-shot learning

Finally, we assess the ability of our models to learn Creoles from few examples.\(^3\) We also compared the results of a pre-trained mBERT, which, unsurprisingly, outperformed all of our checkpoints (corresponding to smaller models learned from tiny data).

---

### Table 2: The hyperparameters used for target Creole and Non-Creole experiments.

<table>
<thead>
<tr>
<th>Hyperparameter</th>
<th>Creole</th>
<th>Non-Creole</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vocabulary size</td>
<td>10,240</td>
<td>10,240</td>
</tr>
<tr>
<td>Learning rate</td>
<td>1.00E-04</td>
<td>5.00E-05</td>
</tr>
<tr>
<td>Weight decay</td>
<td>1.00E-03</td>
<td>1.00E-03</td>
</tr>
<tr>
<td>Dropout</td>
<td>1.00E-01</td>
<td>1.00E-01</td>
</tr>
<tr>
<td>Batch size</td>
<td>256</td>
<td>256</td>
</tr>
</tbody>
</table>

---

\(^{3}\) We also compared the results of a pre-trained mBERT, which, unsurprisingly, outperformed all of our checkpoints (corresponding to smaller models learned from tiny data).
(n=10, ..., 100) at different training stages. Once again, few-shot learning from post-compression checkpoints led to higher perplexity than training from maximum perplexity or early checkpoints.

5 Creoles through the Lens of WALS

We have observed unique patterns for Creoles. Namely, multilingual learning of the related languages did not lead to successful transfer to Creoles; and that Creoles exhibit a unique compression effect. Here, we speculate whether there is a link between these observations, and investigate whether typological features can shed lights into our results. To that effect, we use The World Atlas of Language Structures (WALS)\(^4\), which has been used to study Creoles before (Daval-Markussen and Bakker, 2012). Here, we use the cosine distance between the normalized (full) WALS feature vectors as our distance metric.\(^5\)

In Figure 5, we present an example heatmap for Nigerian Pidgin, which shows that Nigerian Pidgin is less related to ancestor and random languages than any of them internally (except Quechua and Cherokee). We found this pattern present for each of the Creoles. Thus, it would seem that Creoles’ relatively large distance\(^6\) from other languages may make cross-lingual transfer a particular challenge for learning Creoles.\(^7\)

6 Conclusion

We have presented two hypotheses (R1 and R2) about the possibility of zero-shot transfer to Creoles, both built on the idea that Creoles share characteristics with their ancestor languages. This is not exactly equivalent to the so-called superstratist view of Creole genesis, which maintains that Creoles are essentially regional varieties of their European ancestor languages, but if the superstratist view was correct, R1 would very likely be easily validated (Singh et al., 2019). Our results show the opposite trend, however. Zero-shot transfer to Creole languages from their ancestor languages is hard. We do not claim that our results favor an exceptionalist position on Creoles. While we performed a first analysis of several segmentation approaches (i.e., BERT word piece, grapheme-to-phoneme, and byte-pair encodings) – which did not change the training dynamics – we believe that a rigorous comparison would be beneficial for future work in ancestor-to-Creole transfer. We hope that continued investigation in this direction can shed more light on cross-lingual transfer, especially with regards to Creoles, and that this work has demonstrated that not all transfer between related languages is trivial.

7 Acknowledgments

We would like to thank the reviewers for their feedback on this manuscript. This project has received funding from the European Union’s Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 801199 (for Heather Lent and Emanuele Bugliarello) and the Google Research Award (for Heather Lent and Anders Søgaard).

\(^4\)We note that previous work has suggested that WALS features alone may be insufficient for typological comparison of Creoles to non-Creoles (Murawaki, 2016).

\(^5\)We also note that cosine distance might not be meaningful here, as the normalized (full) space does not represent the feature geometry of the space that the linguists that developed the features in WALS were assuming.
References


Benjamin Muller, Antonios Anastasopoulos, Benoît Sagot, and Djamel Seddah. 2021. When being unseen from mBERT is just the beginning: Handling new languages with multilingual language models.


