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.

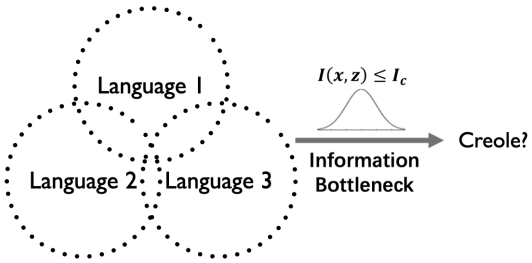

Figure 1: Does the Information Bottleneck principle capture some of the dynamics of Creole formation?

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

Proceedings of the Third Workshop on Insights from Negative Results in NLP, pages 68 - 74

Creole	Ancestors	Random Controls
Nigerian Pidgin	English, Hausa, Yoruba, Portuguese	Afrikaans, Cherokee, Hungarian, Quechua
Jamaican Patois	English, Hausa, Spanish, Igbo	Afrikaans, Cherokee, Hungarian, Quechua
Saint Lucian Creole	French, Hausa, Yoruba, Igbo	Afrikaans, Cherokee, Hungarian, Quechua
Haitian Creole	French, Fon, Spanish, Igbo	Afrikaans, Cherokee, Hungarian, Quechua
Non-Creole	**Relatives**	**Random Controls**
Spanish	French, Italian, Portuguese, Romanian	Afrikaans, Cherokee, Hungarian, Quechua
Danish	Norweigan, Icelandic, Swedish, German	Afrikaans, Cherokee, Hungarian, Quechua

Table 1: Transfer setups in our study. We aim to learn target Creoles and Non-Creoles by training on **1)** their Ancestors or Relatives, respectively; and **2)** languages unrelated to the target ones as a control (Random Controls).

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

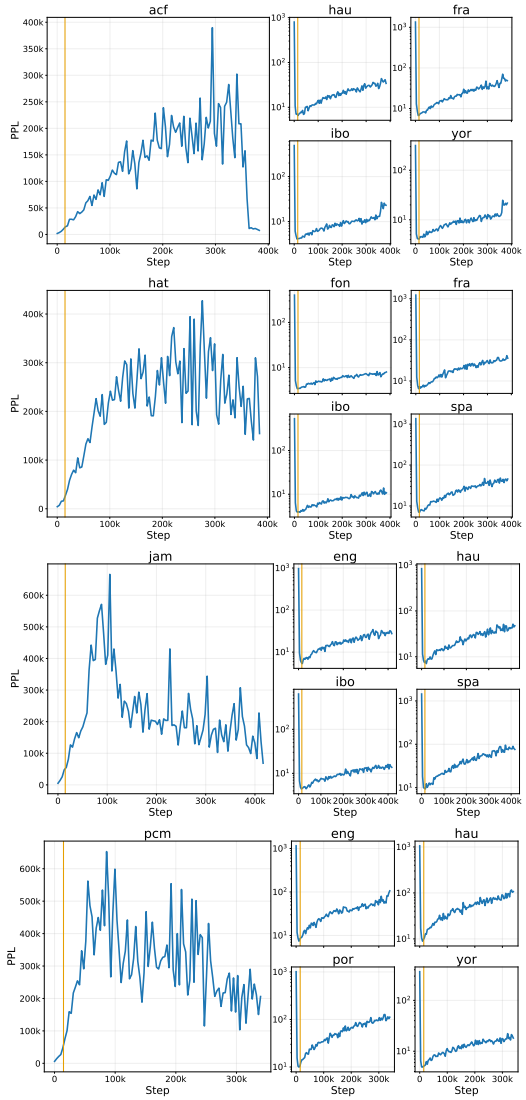

Figure 2: Four zero-shot transfer experiments for Creole languages. The left-hand side plot shows the (zero-shot) validation curve for checkpoints on Creole data; the small plots show the learning curves for the training languages. We see an initial increase in perplexity (disproving **R1**). The yellow 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.

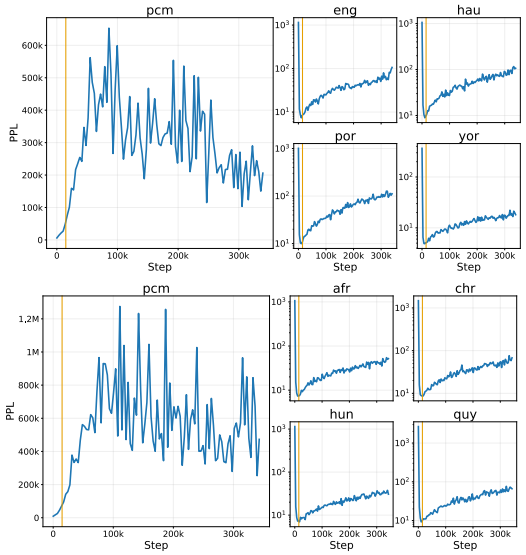

Figure 3: Learning curves for Nigerian Pidgin English when training on **ancestor** languages (top) and when training on **random** languages (bottom). No significant differences are observed. This disproves **R2**.

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

Hyperparameter	Creole	Non-Creole
Vocabulary size	10,240	10,240
Learning rate	1.00E-04	5.00E-05
Weight decay	1.00E-03	1.00E-03
Dropout	1.00E-01	1.00E-01
Batch size	256	256

Table 2: The hyperparameters used for target Creole and Non-Creole experiments. Vocab size, weight decay, and dropout were the same across Creole and Non-Creole experiments, however the Non-Creoles required a smaller learning rate, in order to successfully learn. All experiments were run on a TitanRTX GPU.

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).[1] 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.[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

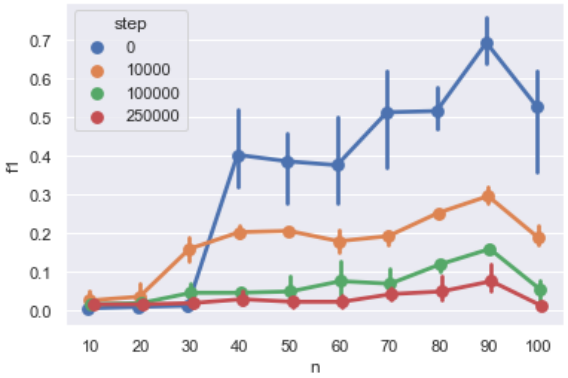

Figure 4: Results for downstream performance on Nigerian Pidgin NER, across 3 random seeds. The top row shows our model trained on ancestor of Nigerian Pidgin (pcm), while the bottom one shows results for mBERT. Step 0 in the legend refers to the pre-trained mBERT, without any further training on ancestor languages.

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 (Adelani 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).

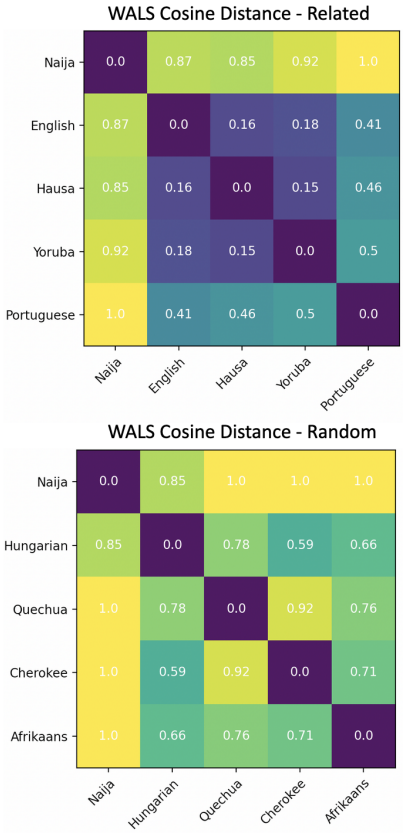

Figure 5: Heatmaps of WALS cosine distances between Nigerian Pidgin (Naija) and its parent and random training languages. We observe that Nigerian Pidgin is *less* related to any of these languages, than any of them internally (except Quechua and Cherokee).

(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] wals.info.

[5] https://github.com/mayhewsw/wals.

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

[7] 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

Enoch Oladé Aboh and Michel DeGraff. 2016. A null theory of creole formation based on universal grammar.

David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D'souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, Stephen Mayhew, Israel Abebe Azime, Shamsuddeen H. Muhammad, Chris Chinenye Emezue, Joyce Nakatumba-Nabende, Perez Ogayo, Aremu Anuoluwapo, Catherine Gitau, Derguene Mbaye, Jesujoba Alabi, Seid Muhie Yimam, Tajuddeen Rabiu Gwadabe, Ignatius Ezeani, Rubungo Andre Niyongabo, Jonathan Mukiibi, Verrah Otiende, Iroro Orife, Davis David, Samba Ngom, Tosin Adewumi, Paul Rayson, Mofetoluwa Adeyemi, Gerald Muriuki, Emmanuel Anebi, Chiamaka Chukwuneke, Nkiruka Odu, Eric Peter Wairagala, Samuel Oyerinde, Clemencia Siro, Tobius Saul Bateesa, Temilola Oloyede, Yvonne Wambui, Victor Akinode, Deborah Nabagereka, Maurice Katusiime, Ayodele Awokoya, Mouhamadane MBOUP, Dibora Gebreyohannes, Henok Tilaye, Kelechi Nwaike, Degaga Wolde, Abdoulaye Faye, Blessing Sibanda, Orevaoghene Ahia, Bonaventure F. P. Dossou, Kelechi Ogueji, Thierno Ibrahima DIOP, Abdoulaye Diallo, Adewale Akinfaderin, Tendai Marengereke, and Salomey Osei. 2021. MasakhaNER: Named entity recognition for African languages. *Transactions of the Association for Computational Linguistics*, 9:1116–1131.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Aymeric Daval-Markussen and Peter Bakker. 2012. Explorations in creole research with phylogenetic tools. In *EACL 2012*.

Michael DeGraff. 2005a. o creole languages constitute an exceptional typological class? *Revue française de linguistique appliquée*, 10(1):11–24.

Michel DeGraff. 2005b. Linguists' most dangerous myth: The fallacy of creole exceptionalism. *Language in Society*, 34:533 – 591.

Ameet Deshpande, Partha Talukdar, and Karthik Narasimhan. 2021. When is bert multilingual? isolating crucial ingredients for cross-lingual transfer.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Philipp Dufter, Martin Schmitt, and Hinrich Schütze. 2020. Increasing learning efficiency of self-attention networks through direct position interactions, learnable temperature, and convoluted attention. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3630–3636, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Philipp Dufter and Hinrich Schütze. 2020. Identifying elements essential for BERT's multilinguality. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4423–4437, Online. Association for Computational Linguistics.

Karthikeyan K, Zihan Wang, Stephen Mayhew, and Dan Roth. 2020. Cross-lingual ability of multilingual bert: An empirical study. *ArXiv*, abs/1912.07840.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *CoRR*, abs/2001.08361.

Phillip Keung, Yichao Lu, Julian Salazar, and Vikas Bhardwaj. 2020. Don't use English dev: On the zero-shot cross-lingual evaluation of contextual embeddings. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 549–554, Online. Association for Computational Linguistics.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot language transfer with multilingual Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499, Online. Association for Computational Linguistics.

Heather Lent, Emanuele Bugliarello, Miryam de Lhoneux, Chen Qiu, and Anders Søgaard. 2021. On language models for creoles. In *Proceedings of the 25th Conference on Computational Natural Language Learning (CoNLL)*, page (to appear), Punta Cana, Dominican Republic. Association for Computational Linguistics.

Thomas Mayer and Michael Cysouw. 2014. Creating a massively parallel Bible corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3158–3163, Reykjavik, Iceland. European Language Resources Association (ELRA).

John H. McWhorter. 1998. Identifying the creole prototype: Vindicating a typological class. 74(4):788–818.

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

In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 448–462, Online. Association for Computational Linguistics.

Yugo Murawaki. 2016. Statistical modeling of creole genesis. In *NAACL*.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2021. UNKs everywhere: Adapting multilingual language models to new scripts. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10186–10203, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Jasdeep Singh, Bryan McCann, Richard Socher, and Caiming Xiong. 2019. BERT is not an interlingua and the bias of tokenization. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 47–55, Hong Kong, China. Association for Computational Linguistics.

Raymond Hendy Susanto, Ohnmar Htun, and Liling Tan. 2019. Sarah's participation in WAT 2019. In *Proceedings of the 6th Workshop on Asian Translation*, pages 152–158, Hong Kong, China. Association for Computational Linguistics.

Naftali Tishby, Fernando C. Pereira, and William Bialek. 1999. The information bottleneck method. pages 368–377.

Rob van der Goot, Ahmet Üstün, Alan Ramponi, Ibrahim Sharaf, and Barbara Plank. 2021. Massive choice, ample tasks (machamp):a toolkit for multi-task learning in nlp.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.