On Language Models for Creoles

Lent, Heather; Bugliarello, Emanuele; De Lhoneux, Miryam; Qiu, Chen; Søgaard, Anders

Published in:
Proceedings of the 25th Conference on Computational Natural Language Learning

DOI:
10.18653/v1/2021.conll-1.5

Publication date:
2021

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

Document license:
CC BY

Citation for published version (APA):
On Language Models for Creoles
Heather Lent\textsuperscript{c} Emanuele Bugliarello\textsuperscript{c} Miryam de Lhoneux\textsuperscript{c,\textsuperscript{o,\textsuperscript{o}}} Chen Qiu\textsuperscript{u} Anders Søgaard\textsuperscript{b}
\textsuperscript{c} University of Copenhagen, Denmark
\textsuperscript{o} Uppsala University, Sweden \textsuperscript{b} KU Leuven, Belgium
\textsuperscript{u} Wuhan University of Science and Technology, China
\{hcl, emanuele, ml, soegaard\}@di.ku.dk
chen@wust.edu.cn

Abstract
Creole languages such as Nigerian Pidgin English and Haitian Creole are under-resourced and largely ignored in the NLP literature. Creoles typically result from the fusion of a foreign language with multiple local languages, and what grammatical and lexical features are transferred to the creole is a complex process (Sessarego, 2020). While creoles are generally stable, the prominence of some features may be much stronger with certain demographics or in some linguistic situations (Winford, 1999; Patrick, 1999). This paper makes several contributions: We collect existing corpora and release models for Haitian Creole, Nigerian Pidgin English, and Singaporean Colloquial English. We evaluate these models on intrinsic and extrinsic tasks. Motivated by the above literature, we compare standard language models with distributionally robust ones and find that, somewhat surprisingly, the standard language models are superior to the distributionally robust ones. We investigate whether this is an effect of over-parameterization or relative distributional stability, and find that the difference persists in the absence of over-parameterization, and that drift is limited, confirming the relative stability of creole languages.

1 Introduction
A creole language arises if a pidgin,\textsuperscript{1} developed by adults for use as a second language, becomes the native and primary language of their children. Although a large portion of creole languages have their roots in Western European colonialism and slavery, creole languages still serve as important lingua franca in multi-ethnic and multilingual communities, and creoles are often an important part of the local identity. Moreover, there are more than a hundred million speakers of creole languages worldwide (Fig. 1), with similar needs for technological assistance, and yet creoles are still largely absent from NLP research (Joshi et al., 2020). Haitian Creole, for example, has 9.6 million speakers as of today; Nigerian Pidgin English has 100 million speakers, and Singaporean Colloquial English (Singlish) has 3.5 million speakers. This paper sets out to collect existing resources for these three languages and provides language models for them. In doing so, we wish to take the nature of creole languages into account, not necessarily assuming that our best approaches to modeling non-creole language are also best for the creole languages.

The nature of creole languages has been a matter of much debate in linguistics during the last decade (Sessarego, 2020): Some see creole languages as natural stages in language change cycles (Aboh, 2015), while others see them as a distinct typological class with unique characteristics, including, for example, a very simple morphology (McWhorter, 1998). Another feature of creoles is that they exhibit significant variation across groups of speakers (Patrick, 1999). Winford (1999) goes as far as to call creoles a continua that cannot be captured under a single grammar.

Consider the following pair of sentences from

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{map.png}
\caption{Creoles with a minimum of a hundred thousand speakers are shown here (Hawaiian Pidgin not pictured). Approximately 180 million creole speakers are represented in this map. Data extracted from https://en.wikipedia.org/wiki/List_of_creole_languages.}
\end{figure}
Bajpai et al. (2017):

(1) John sibei hum sup one.

(2) John very buaya sia.

Here, according to the authors, both sentences are valid utterances in Singlish, and they both mean John is so lecherous, but the first would more likely come from a speaker of Chinese, and the second from a Malay speaker. From this, we derive the conjecture that creole language models can benefit from learned mixtures of source languages. Training on mixtures of source languages has been applied to language modeling of code-switched language (Pratapa et al., 2018), and it is clear from examples such as the one in Fig. 2 that creole languages, at the sentence level, share commonalities with code-switched language, with vocabularies drawn from multiple source languages. To exploit synergies with learned mixtures of source languages, and to obtain robust performance across related, but unseen distributions, we explore ways of training creole language models with distributionally robust objectives (Oren et al., 2019). Our results below, however, show that, somewhat surprisingly, this conjecture is probably not true, at least not in a straightforward way.

Contributions  We combine existing datasets and present pretrained language models for the following creole languages: Nigerian Pidgin English, Singaporean Colloquial English (Singlish), and Haitian Creole. We perform intrinsic evaluation (word prediction), as well as extrinsic evaluation (part-of-speech tagging and named entity recognition). Comparing language models trained with empirical risk minimization to languages models trained with robust objectives, we observe that training with multiple related languages does not improve creole modeling; and also, somewhat surprisingly, that models with empirical risk minimization are superior to models robust across domains. We hence investigate why this is: in particular, whether it is due to over-parameterization, insufficient regularization (Sagawa et al., 2019), or relative distributional stability (Ben-David et al., 2007). We observe no significant difference for language models with fewer parameters or higher degree of regularization. On the other hand, we find that the underlying reason might be the relative stability of the creoles, which show no significant drift.

2 Related Work

NLP research on creoles  Despite the unique features of creoles that make them an interesting application for multilingual and cross-lingual NLP, as well as the open-ended debate about the linguistic nature of creoles (Sessarego, 2020), little attention has been devoted to creoles in NLP. (We present the works related to the specific creoles of focus in this paper in §3.) One relevant work by Murawki (2016) explored the typological status of creoles and also introduced a method for statistical modeling of creole genesis. To start, the authors reported that binary SVM classification of creole and non-creole languages failed to distinguish the two classes, even though their underlying distributions are quite different. After this, they introduce a statistical model of creoles, formulated as a mixture of its influential languages and an inferred “restructuring”, which is set of possible linguistic feature distributions that are observed across languages included in their experiments. Overall, this work showcases how statistical modeling methods can be useful for investigating the language evolution of creoles, however there is also no discussion of how their findings could help others extend current NLP methods for creoles.
NLP research on pidgins and code-switching

Creoles are pidgins that have consolidated over time to become a first language for new generations of speakers. The NLP literature on pidgins is even more sparse than the literature on creoles, because many pidgins that did not undergo creolization have gone extinct, such as Maritime Polynesian Pidgin (Kriegel, 2016). Code-switching literature, however, is also relevant, as both pidgins and creoles also draw from other languages. Importantly, pidgins differ from code-switching or mixed language in that code-switching typically only occurs between two bilingual or highly proficient speakers of two languages. Pidgins, on the other hand, are derived from multiple languages, and spoken by those who do not fluently speak every language involved. The NLP literature on code-switching is surprisingly rich, however. We refer readers to Çetinoğlu et al. (2016) and Doğruöz et al. (2021) for an overview.

Computational research on language evolution

Research on creoles is more common in the field of language evolution than in NLP. In particular, work on creoles in this field typically focuses on their computational modeling, their emergence (Nakamura et al., 2009), and their evolution (Jansson et al., 2015; Furman and Nitschke, 2020). Other creole modeling efforts in this space may be more tailored towards specific linguistic insights (Parkvall, 2008). While these studies demonstrate that work on creoles is being done in a computational space, it is difficult to apply conclusions from them to NLP, because distinct empirical assumptions are made in these two research areas.

Distributionally robust optimization

Effectively learning to model and predict underrepresented subdistributions has always been a challenge in machine learning, e.g., when predicting rare classes, (Scheirer et al., 2013; Fei and Liu, 2016) or classes of examples from rare domains (Zheng et al., 2020) or minority groups (Hashimoto et al., 2018). Often, underrepresented data is ignored or learned poorly by the models (Feldman and Zhang, 2020), compared to their over-represented counterparts. Distributionally Robust Optimization (DRO) (Hashimoto et al., 2018; Sagawa et al., 2019) aims to minimize the loss on all sub-populations, rather than minimizing their average (Ben-Tal et al., 2013). DRO has been particularly useful in the domain of algorithmic fairness (Hashimoto et al., 2018), but has also been found to boost performance on underrepresented domains in language modeling (Oren et al., 2019) and is generally applicable in situations with drift (Koh et al., 2021).

3 Creoles and Corpora

While creole languages are spoken by hundreds of millions, and are often a lingua franca within a larger community, only a handful of resources exist for creoles presently. Some challenges to collecting data resources for creole languages can be a creole’s non-standardized orthography, e.g. Haitian Creole (Hewavitharana et al., 2011), or the specific contexts in which creoles are used – it may not always be used in official capacities for news, education, and official documents, even if the creoles are widely used in most other aspects of life (Shah-Sanghavi, 2017). This of course complicates data collection. In this work, we focus on the following creoles, as they each have diverse linguistic makeup and have some existing datasets:

Nigerian Pidgin English

West Africa is one of the world’s most linguistically diverse places, with Nigeria alone having over 400 languages (Ufomata, 1999). Recent work to advance African NLP has led to the creation of several datasets in Nigerian Pidgin English (Agić and Vulić, 2019; Ogueji and Ahia, 2019; Ndubuisi-Obi et al., 2019; Caron et al., 2019; Oyewusi et al., 2021; Adelani et al., 2021; Oyewusi et al., 2020), which makes it particularly well-resourced in comparison to other creole languages. Nigerian Pidgin English, also referred to as simply Nigerian Pidgin, can further be understood as a member in the larger family of West African Pidgins, as many West African countries have their own unique variation of this creole, but all share influences from many of the same languages, such as Igbo, Hausa, and Yoruba.

The first sizeable Nigerian Pidgin dataset comes from Agić and Vulić (2019), who collected parallel text from several magazines written by a religious society, which have parallel translations in many languages. This dataset has been utilized in the first attempts to develop baselines for machine translation of Nigerian Pidgin English (Ogueji and Ahia, 2019; Ahia and Ogueji, 2020). Furthermore, Ogueji and Ahia (2019) also introduced the first corpus of Nigerian Pidgin English to further facilitate machine translation of Nigerian Pidgin into English. Ndubuisi-Obi et al. (2019) also introduced a code-switching corpus of news articles and on-
line comments in both Nigerian Standard English and Nigerian Pidgin. In this work, they discuss some challenges of working with Nigerian Pidgin, such as non-standardized spelling. They also find that different topics prompt code-switching to Nigerian Pidgin over Nigerian Standard English. More task-specific Nigerian Pidgin datasets have been introduced for Universal Dependency Parsing (Caron et al., 2019), named entity recognition (Oyewusi et al., 2021; Adelani et al., 2021), sentiment analysis (Oyewusi et al., 2020), and speech recognition (Bigi et al., 2017; Ajisafe et al., 2020).

**Singlish** Singaporean Colloquial English, also known as Singlish, has English as a source language, but also draws parts of its grammar and vocabulary from languages such as Mandarin, Cantonese, Hakka, Hokkien, Malay, and Tamil. Presently, few publicly available datasets exist in Singlish, as this creole is primarily utilized for informal conversation between people and not for official purposes. The largest relevant corpus is The National University of Singapore SMS Corpus from Chen and Min-Yen (2015), which consists of over 67,000 text messages written by Singaporeans. Qualitatively, we observed that this dataset is much closer to Standard English, albeit with noise from outdated SMS language, than the example provided in Fig. 2, but, within this data, we still observe many hallmark features of Singlish such as discourse markers and vocabulary from relevant languages. Tan et al. (2020) have also released a webcrawler that collects posts from an popular Singaporean forum about hardware, where discussion is often in Singlish. They use the resulting Singlish corpus as part of their work to investigate the role of inflection for NLP with non-standard forms of English. Beyond plain text corpora, Wang et al. (2017) introduced the first Singlish Universal Dependency dataset, which was further expanded upon in Wang et al. (2019). Chau et al. (2020) used this dataset as a low-resource language test case for their method of pretraining mBERT (Devlin et al., 2019). Finally, a few studies have been done on private datasets for sentiment analysis (Bajpai et al., 2017; Ho et al., 2018), and polarity detection (Lo et al., 2016).

**Haitian Creole** Haitian Creole exhibits a combination of French with many West African languages (e.g. Igbo, Yoruba, Fon, etc.). Haitian Creole seized the attention of the machine translation community in the aftermath of the 2010 earthquake crisis in Haiti, during which Munro (2010, 2013) developed the Haitian Disaster Response Corpus. This is a parallel Haitian–English dataset of SMS messages related to the crisis, to enable rapid development of machine translation systems to assist the crisis response. This dataset was included in the 2011 Workshop for Machine Translation (Callison-Burch et al., 2011), in conjunction with data from the medical domain, newswire, and a Haitian glossary.\(^3\) Several studies used this data to extend methods in statistical machine translation (Hu et al., 2011a,b; R. Costa-jussá and Banchs, 2011) as well as spell checking and data cleaning (Stymne, 2011).

### Datasets for Creole Language Models

We experiment with training language models for creoles with a mixture of creole data, and additional data from languages influential to each creole.

<table>
<thead>
<tr>
<th>Language</th>
<th>Source</th>
<th>Domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>en, fr, es, pt, yo, zh, ta</td>
<td>WMT-News 2020</td>
<td>news</td>
</tr>
<tr>
<td>ms</td>
<td>Malay 30k News</td>
<td>news</td>
</tr>
<tr>
<td>Nigerian Pidgin</td>
<td>PidginUNMT Corpus</td>
<td>news</td>
</tr>
<tr>
<td>Singlish</td>
<td>Singapore SMS Corpus</td>
<td>sms</td>
</tr>
<tr>
<td>Haitian Creole</td>
<td>Disaster Response Corpus</td>
<td>sms</td>
</tr>
</tbody>
</table>

Table 1: Data resources utilized in our experiments.

<table>
<thead>
<tr>
<th>Creole</th>
<th>Langs</th>
<th># Train Mixed-Lang</th>
<th># Train Creole-Only</th>
<th># Dev Creole-Only</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nigerian Pidgin</td>
<td>en, pt, yo</td>
<td>230,105</td>
<td>53,006</td>
<td>3,359</td>
</tr>
<tr>
<td>Singlish</td>
<td>en, zh, ms, ta</td>
<td>265,030</td>
<td>67,615</td>
<td>2,790</td>
</tr>
<tr>
<td>Haitian Creole</td>
<td>fr, yo, es</td>
<td>32,768</td>
<td>8,192</td>
<td>988</td>
</tr>
</tbody>
</table>

Table 2: Creoles, their influential languages (Langs), and the number of examples in the Train-Dev split for our MIXED-LANGUAGE and CREOLE-ONLY experiments. Both use the same creole-only dev dataset.

---

\(^3\)http://www.speech.cs.cmu.edu/haitian/text/.
formation about the original language of each sentence, so that we can form language-specific groups for DRO (see §5.1 for more details on DRO grouping). The total number of train and development examples were determined by the number of sentences in the base (creole) dataset for a 95-5 train-development split. Singlish had equal representation of each language, with 53,006 examples per language, including Singlish. Haitian Creole also had equally represented languages, with 8,192 examples for Haitian and each additional language. For the Nigerian Pidgin MIXED-LANGUAGE dataset, English, Portuguese, and Nigerian Pidgin were composed equally with 67,615 examples each, and Yoruba with only 27,260 examples due to the small size of the original data. Thus, we included 95% of the Yoruba WMT-News 2020 dataset.

Language identification within creoles As we will see in §6, training the language models on the MIXED-LANGUAGE dataset with DRO fails to produce positive results. Following from this, we also create a CREOLE-ONLY dataset, composed of only the creole examples. In order to sort the creole examples into distinct groups for DRO, we label each creole example by the collection of the selected languages present in the sentences, as determined by a language identification algorithm. Consider the following examples from their respective CREOLE-ONLY datasets:

Singlish: "treat him makah lah"
en: 88.19%, ms: 4.34%, ta: 0.04%, and zh: 0.01%

Nigerian Pidgin: "Pikin wey like to play wit wetin no dey common and sabi one particular subject reach ground"
en: 87.46%, pt: 0.23%, and yo: 0.03%

Haitian Creole: "Infomation sou kestion te tranble a ak lekol"
fr: 3.50%, es: 0.08%, and yo: 0.01%

While the language identification algorithm is not perfect, the confidence scores for the languages still reflect the high-level trends for the creole examples, namely, that English and Malay ("makan") are indeed present in the Singlish sample, and also that English and Portuguese ("pikin", "sabi") are present in the Nigerian Pidgin example. However, for the Haitian Creole example, we see that none of our chosen languages have very high scores from the language identification algorithm, which begs the question: were there other languages with higher confidence from the language identification algorithm?

To ensure that the languages we chose are well-represented in the creole examples, we looked at the distribution of the identified languages across examples in our CREOLE-ONLY datasets in Fig. 3. From this, we observe that choosing to identify languages specifically related to the creole (i.e. the same languages we included in the MIXED-LANGUAGE datasets) is more reliable than trusting the language identifier pick the top five languages with the highest confidence – there appears to be a bias for falsely predicting European languages, even on creole data unrelated to these languages, as well as some strange outliers, such as Tagalog being the third most commonly predicted language for Haitian Creole sentences. Also, we see that Haitian Creole itself was a commonly identified language, which could explain the low confidence scores for French and Spanish in the example above. Finally, among our specifically chosen languages for the creoles, we see that, although the source language (e.g. English or French) is most dominant, the other languages are still well distributed, with the exception of Yoruba. We surmise that the densely distributed, low-confidence scores for Yoruba can probably be attributed to the fact that Yoruba is a lower-resourced language.

---

https://fasttext.cc/blog/2017/10/02/blog-post.html.
5 Experiments

In this section, we detail our experimental setups. We make our code and models publicly available.

5.1 Training

Using the datasets described above, we conduct several experiments to assess how different training strategies affect the modeling of creoles. We conduct all the experiments on both English BERT and multilingual mBERT models (Devlin et al., 2019). As our baseline, we consider pretrained BERT and mBERT models, and evaluate them on our development splits for the creoles. We then assess the effectiveness of two popular training strategies: Empirical Risk Minimization (ERM) and Distributionally Robust Optimization (DRO). In this case, ERM consists of masked language modeling over all the data points in each dataset, in a similar fashion as done during pretraining.

For DRO, we utilize the WILDS library (Koh et al., 2020), which uses metadata associated with the input data to form the groups for DRO. In our case, we investigate three grouping strategies: grouping with language information as metadata (DRO-Language), as well as with two additional control experiments. In the first control experiment, we assign all training examples to the same group (DRO-One), such that that DRO is optimizing over only one large group. In the second control experiment, we randomly assign examples to one of four groups (DRO-Random). The motivation for these control experiments is to ensure that improvements for DRO are actually grounded in the language information, and not an artifact of the WILDS grouping algorithm.

In DRO-Language, information about the examples’ language makeup is used to determine the groups. In Mixed-LANGUAGE, we rely on our knowledge of where the examples were sampled from, but in Creole-ONLY, we subdivide the creole examples depending on their etymology. Specifically, grouping is done as follows in our two data setups outlined in §4:

• Mixed-LANGUAGE: Here, grouping is done over the languages in the training data. For example, in the case of Nigerian Pidgin, if a sentence originally comes from the Yoruba corpus, it is assigned to the Yoruba group, and similarly for Nigerian Pidgin and the other languages listed in Tab. 2 for each creole.

• Creole-ONLY: Here, as we only have the creole samples, grouping is done over the confidence scores from the collection of the influential languages (see Section §4). An example is assigned to one of $2^N$ groups, representing the combinations of detected languages in a sentence. $N$ is the number of languages listed in Tab. 2 (Langs) for each creole, and presence of a language is derived from its confidence score by the language identifier: if there is a confidence of 0.1% or higher that the language is represented in the sentence, then it is considered as present.

5.2 Evaluation

We perform two types of evaluation: intrinsic – based on the MLM training objective – and extrinsic – on traditional downstream NLP tasks.

Intrinsic evaluation We evaluate our language models intrinsically with the following metrics:

• Precision at $k$ ($P@k$): Precision of the language model in predicting a random masked token per sentence. This allows us to assess the general performance following the training objective. In the following, we report P@1. Results at $k = \{5, 10\}$ are in the App.

• Dictionary-based precision at $k$ ($P_D@k$): Due to their nature, most of the words in a creole sentence are from the corresponding source language (see Fig. 3). Hence, for a more principled measurement of precision, we collect online dictionaries of our creoles.

We perform the same MLM task as above, but this time only mask words belonging to the creole dictionaries. By doing so, we can obtain a more accurate measure of what the LMs have learned. We again report results at $k = 1$ here, and refer the reader to the App. for $k = \{5, 10\}$.

• Mean pseudo-log-likelihood score (PLL): Following recent studies (Shin et al., 2019; Wang and Cho, 2019; Salazar et al., 2020), we measure the pseudo-log-likelihood scores from MLMs given by summing the conditional log probabilities $\log P_{MLM}(w_t|w_{<t})$ of each token $w_t$ in a sentence $w = \langle w_1, \ldots, w_T \rangle$. These are obtained in BERT

---

5https://github.com/hclent/creole-dro

6Nigerian Pidgin: http://naijalingo.com/
Singlish: http://www.mysmu.edu/faculty/jacklee/
Table 3: Intrinsic evaluation: Precision@1 (P@1), Precision@1 for words in our creole dictionary (P\textsubscript{D}@1), and average Pseudo-log-likelihood score (PLL). We report results for MIXED-LANGUAGE (top) and CREOLE-ONLY (bottom). We note that ERM consistently outperforms the language models trained with robust objectives.

\begin{table}[h]
\centering
\begin{tabular}{|c|c|c|c|c|c|c|c|c|}
\hline
 & \multicolumn{3}{c|}{Nigerian Pidgin} & \multicolumn{3}{c|}{Singlish} & \multicolumn{2}{c|}{Haitian Creole} \\
\hline
 & P@1 & P\textsubscript{D}@1 & PLL & P@1 & P\textsubscript{D}@1 & PLL & P@1 & P\textsubscript{D}@1 \\
\hline
Pretrained & 22.79 & 10.92 & 142.65 & 23.94 & 21.09 & 76.01 & 18.84 & 5.65 & 177.40 \\
\hline
ERM & 63.83 & 59.97 & 42.41 & 46.77 & 42.89 & 41.06 & 68.09 & 43.35 & 55.04 \\
\hline
DRO-One & 60.99 & 56.76 & 52.51 & 44.23 & 40.73 & 49.18 & 57.04 & 36.73 & 121.51 \\
DRO-Random & 60.40 & 53.36 & 52.69 & 43.33 & 39.07 & 49.14 & 57.65 & 36.16 & 119.17 \\
DRO-Language & 60.40 & 54.80 & 54.17 & 43.19 & 39.57 & 48.88 & 55.54 & 36.54 & 118.85 \\
\hline
C-ONLY & 73.72 & 71.38 & 28.14 & 53.80 & 51.26 & 34.22 & 73.15 & 55.50 & 55.51 \\
\hline
DRO-One & 64.28 & 59.86 & 61.81 & 45.34 & 43.59 & 66.53 & 58.16 & 36.91 & 144.46 \\
DRO-Random & 63.72 & 59.31 & 60.31 & 45.73 & 42.40 & 64.16 & 57.65 & 37.41 & 142.04 \\
DRO-Language & 63.58 & 59.74 & 56.82 & 44.73 & 40.57 & 53.72 & 56.94 & 35.50 & 138.60 \\
\hline
\end{tabular}
\caption{Intrinsic evaluation: Precision@1 (P@1), Precision@1 for words in our creole dictionary (P\textsubscript{D}@1), and average Pseudo-log-likelihood score (PLL). We report results for MIXED-LANGUAGE (top) and CREOLE-ONLY (bottom). We note that ERM consistently outperforms the language models trained with robust objectives.}
\end{table}

by replacing \(w_t\) with the special \([\text{MASK}]\) token. Here, we report the mean score given by:

\[
\text{PLL} = \frac{1}{|C|} \sum_{w \in C} \frac{1}{|w|} \sum_{w_t \in w} \log P_{\text{MLM}}(w_t | w_t; \theta),
\]

where \(C\) denotes the evaluation corpus, and \(\theta\) denotes a model’s parameters.

Extrinsic evaluation We also perform an extrinsic evaluation of our models on downstream tasks, for the datasets that are available. Specifically, we train and evaluate models for Nigerian Pidgin NER and POS tagging with Universal Dependencies (Nivre et al., 2020, UPOS), as well as Singlish UPOS. We fine-tune our pretrained language models on the training sets of these two tasks and evaluate them on the corresponding test sets.

5.3 Framework
We write our code in PyTorch (Paszke et al., 2019). In particular, for language model training, we rely on the HuggingFace Transformers library (Wolf et al., 2019), and the WILDS library (Koh et al., 2020) for DRO. Models are fine-tuned for 100,000 steps with batch size of 16. For downstream tasks, we use MaChAmp (van der Goot et al., 2021) and train our models for 10 epochs. The best check-points were selected based on performance on the dev sets. Unless otherwise specified, we use the default hyperparameters. Our experiments are run on one NVIDIA TitanX GPU in a shared cluster.

6 Results and Analyses

Extrinsic evaluation For all grouping strategies across all metrics. We also observe that \(P_{D}@k\) is a more difficult task than the standard precision at \(k\), with randomly masked tokens (see App. §A for full results with both BERT and mBERT). Moreover we find that the DRO models often have a much higher perplexity than ERM. Finally, the results show that, between the MIXED-LANGUAGE and CREOLE-ONLY experiments, the latter performed better, demonstrating that training on additional data was not useful for learning language models for creoles. While we only report results for BERT here, we observe the same patterns with mBERT (see App. §A).

Extrinsic evaluation Here, we observe the same trend as in the intrinsic evaluation: ERM performs better than DRO (see Tab. 4). Although for Nigerian Pidgin DRO-Language performs better than ERM on both NER and UPOS, the gap between the scores is too small to draw concrete conclusions from.

There are several factors that could have influenced the DRO models to perform worse than ERM. We explore their effects below.
Over-parameterization

Over-parameterization is known to be problematic for DRO (Sagawa et al., 2019). In order to investigate the role of over-parameterization in our experiments, we ran additional MIXED-LANGUAGE experiments on Nigerian Pidgin English, with different sized BERT models, namely BERTTiny, BERTSmall (Jiao et al., 2020), and BERTBase. The results in Tab. 5 demonstrate that over-parameterization was not a leading cause for DRO failure, otherwise we would expect for smaller BERT versions to have relative better performance compared to the corresponding ERM runs. Instead, we see that standard BERT works fine for this task, and over-parameterization is not the cause of poor performance of DRO in our experiments.

Regularization

Sagawa et al. (2019) also discuss how lack of regularization lead to problems for DRO, and how increased regularization is necessary for worst-group generalization. To investigate this potential weakness in our experiments, we run additional experiments using BERTSmall on MIXED-LANGUAGE data for Nigerian Pidgin English, trying different weight decay values in each Tab. 6. If our DRO models were suffering from insufficient regularization, we would expect that increasing the regularization factor of weight decay would boost performance. However, we find no meaningful effect of this hyperparameter, which leads us to believe that insufficient regularization is not a driving factor in the underperformance of DRO compared to ERM.

Drift and creole stability

Creole languages arise from pidgins, which are initially developed for use as second language. Recent years have seen renewed interest in the classic question of the relationship between pidgin and creole formation and second language acquisition (Plag, 2009). To investigate the matter of creole stability, we follow (Ben-David et al., 2007) and calculate the proxy $A$-distance (PAD) between different domains of creole data (see Tab. 7). Specifically, we train an SVM on the BERT encodings. Our $A$-distance results suggest that creole languages do not exhibit more drift than English when the data are comparable. This potentially explains why distributionally robust language models do not outperform regular language models trained with empirical risk minimization objectives.

7 Conclusion

In this paper, we bring creole languages to the attention of the NLP community. We collect data and train baseline language models for three creoles, and evaluate these models across the downstream tasks of part-of-speech tagging and named entity recognition. Based on previous work suggesting the instability of creole languages (Winford, 1999; Patrick, 1999), we explore the impact of using more robust learning objectives for masked language modeling of creoles, but our results show that vanilla empirical risk minimization is superior. We show that this is not the result of over-parameterization or lack of regularization, but in-
stead suggest this is a result of the relative stability of creole languages. We note that it still remains possible that significant improvements could be achieved by modeling dynamics specific to creole languages, i.e., the processes that govern their development, including social factors (Holm, 2000) and second language acquisition dynamics (Plag, 2009).

8 Acknowledgments

We would like to thank Robert Monarch and Chris Callison-Burch for their help with the Haitian data. 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), the Swedish Research Council Grant 2020-00437 (for Miryam de Lhonneux), and the Google Research Award (for Heather Lent and Anders Søgaard).

References


B. Bigi, B. Caron, and Oyelere S. Abiola. 2017. Developing resources for automated speech processing of the african language naija (Nigerian pidgin).


S. Sessarego. 2020. Not all grammatical features are robustly transmitted during the emergence of creoles. *Humanities and Social Sciences Communications*, 7:1–8.


### Appendix

#### Singlish

<table>
<thead>
<tr>
<th>Model</th>
<th>Strategy</th>
<th>P@1</th>
<th>P@5</th>
<th>P@10</th>
<th>PLL</th>
<th>P_D@1</th>
<th>P_D@5</th>
<th>P_D@10</th>
</tr>
</thead>
<tbody>
<tr>
<td>BERT</td>
<td>ERM</td>
<td>46.77</td>
<td>68.89</td>
<td>74.34</td>
<td>41.07</td>
<td>42.89</td>
<td>66.76</td>
<td>74.17</td>
</tr>
<tr>
<td></td>
<td>DRO-One</td>
<td>44.23</td>
<td>64.73</td>
<td>71.90</td>
<td>49.18</td>
<td>40.73</td>
<td>63.05</td>
<td>70.13</td>
</tr>
<tr>
<td></td>
<td>DRO-Random</td>
<td>43.33</td>
<td>65.63</td>
<td>71.58</td>
<td>49.14</td>
<td>39.07</td>
<td>61.02</td>
<td>68.42</td>
</tr>
<tr>
<td></td>
<td>DRO-Language</td>
<td>43.19</td>
<td>64.80</td>
<td>71.22</td>
<td>48.88</td>
<td>39.57</td>
<td>61.54</td>
<td>70.34</td>
</tr>
<tr>
<td>mBERT</td>
<td>ERM</td>
<td>47.78</td>
<td>68.82</td>
<td>75.77</td>
<td>42.26</td>
<td>37.00</td>
<td>61.71</td>
<td>70.85</td>
</tr>
<tr>
<td></td>
<td>DRO-One</td>
<td>44.37</td>
<td>65.13</td>
<td>72.54</td>
<td>50.99</td>
<td>33.71</td>
<td>57.79</td>
<td>65.72</td>
</tr>
<tr>
<td></td>
<td>DRO-Random</td>
<td>44.34</td>
<td>65.95</td>
<td>72.44</td>
<td>50.39</td>
<td>35.25</td>
<td>59.20</td>
<td>66.93</td>
</tr>
<tr>
<td></td>
<td>DRO-Language</td>
<td>43.69</td>
<td>64.91</td>
<td>71.61</td>
<td>50.49</td>
<td>33.45</td>
<td>58.91</td>
<td>67.42</td>
</tr>
</tbody>
</table>

#### Naija

<table>
<thead>
<tr>
<th>Model</th>
<th>Strategy</th>
<th>P@1</th>
<th>P@5</th>
<th>P@10</th>
<th>PLL</th>
<th>P_D@1</th>
<th>P_D@5</th>
<th>P_D@10</th>
</tr>
</thead>
<tbody>
<tr>
<td>BERT</td>
<td>ERM</td>
<td>63.83</td>
<td>80.52</td>
<td>85.44</td>
<td>42.41</td>
<td>59.97</td>
<td>78.72</td>
<td>83.93</td>
</tr>
<tr>
<td></td>
<td>DRO-One</td>
<td>60.99</td>
<td>77.52</td>
<td>82.94</td>
<td>52.51</td>
<td>56.76</td>
<td>76.21</td>
<td>81.94</td>
</tr>
<tr>
<td></td>
<td>DRO-Random</td>
<td>60.40</td>
<td>78.44</td>
<td>82.88</td>
<td>52.69</td>
<td>56.33</td>
<td>75.27</td>
<td>81.17</td>
</tr>
<tr>
<td></td>
<td>DRO-Language</td>
<td>60.40</td>
<td>77.40</td>
<td>82.69</td>
<td>54.18</td>
<td>54.80</td>
<td>74.48</td>
<td>80.32</td>
</tr>
<tr>
<td>mBERT</td>
<td>ERM</td>
<td>62.68</td>
<td>80.52</td>
<td>85.55</td>
<td>44.98</td>
<td>62.19</td>
<td>82.25</td>
<td>87.08</td>
</tr>
<tr>
<td></td>
<td>DRO-One</td>
<td>60.76</td>
<td>77.74</td>
<td>82.88</td>
<td>58.15</td>
<td>56.43</td>
<td>77.61</td>
<td>82.67</td>
</tr>
<tr>
<td></td>
<td>DRO-Random</td>
<td>60.34</td>
<td>77.23</td>
<td>82.24</td>
<td>57.90</td>
<td>56.70</td>
<td>77.50</td>
<td>82.51</td>
</tr>
<tr>
<td></td>
<td>DRO-Language</td>
<td>58.88</td>
<td>76.56</td>
<td>81.73</td>
<td>59.91</td>
<td>54.99</td>
<td>76.63</td>
<td>82.50</td>
</tr>
</tbody>
</table>

#### Haitian

<table>
<thead>
<tr>
<th>Model</th>
<th>Strategy</th>
<th>P@1</th>
<th>P@5</th>
<th>P@10</th>
<th>PLL</th>
<th>P_D@1</th>
<th>P_D@5</th>
<th>P_D@10</th>
</tr>
</thead>
<tbody>
<tr>
<td>BERT</td>
<td>ERM</td>
<td>68.09</td>
<td>82.98</td>
<td>87.34</td>
<td>55.05</td>
<td>43.35</td>
<td>63.89</td>
<td>71.35</td>
</tr>
<tr>
<td></td>
<td>DRO-One</td>
<td>57.04</td>
<td>71.12</td>
<td>75.58</td>
<td>121.51</td>
<td>36.73</td>
<td>52.55</td>
<td>58.25</td>
</tr>
<tr>
<td></td>
<td>DRO-Random</td>
<td>57.65</td>
<td>71.53</td>
<td>75.79</td>
<td>119.17</td>
<td>36.16</td>
<td>50.63</td>
<td>56.38</td>
</tr>
<tr>
<td></td>
<td>DRO-Language</td>
<td>57.55</td>
<td>71.23</td>
<td>75.28</td>
<td>118.85</td>
<td>36.69</td>
<td>50.48</td>
<td>55.89</td>
</tr>
<tr>
<td>mBERT</td>
<td>ERM</td>
<td>60.79</td>
<td>76.70</td>
<td>81.56</td>
<td>60.27</td>
<td>46.35</td>
<td>64.56</td>
<td>70.96</td>
</tr>
<tr>
<td></td>
<td>DRO-One</td>
<td>51.06</td>
<td>65.45</td>
<td>69.71</td>
<td>148.12</td>
<td>34.57</td>
<td>49.30</td>
<td>55.61</td>
</tr>
<tr>
<td></td>
<td>DRO-Random</td>
<td>50.86</td>
<td>65.05</td>
<td>69.50</td>
<td>146.18</td>
<td>34.52</td>
<td>49.58</td>
<td>55.00</td>
</tr>
<tr>
<td></td>
<td>DRO-Language</td>
<td>50.15</td>
<td>64.54</td>
<td>69.40</td>
<td>145.97</td>
<td>33.55</td>
<td>48.21</td>
<td>55.08</td>
</tr>
</tbody>
</table>

Table 8: Full results for Mixed-Language experiments.
### Table 9: Full results for Creole-Only experiments.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Model</th>
<th>Strategy</th>
<th>P@1</th>
<th>P@5</th>
<th>P@10</th>
<th>PLL</th>
<th>P_D@1</th>
<th>P_D@5</th>
<th>P_D@10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Singlish</td>
<td>BERT</td>
<td>ERM</td>
<td>53.80</td>
<td>75.02</td>
<td>80.36</td>
<td>34.22</td>
<td>51.26</td>
<td>74.09</td>
<td>80.15</td>
</tr>
<tr>
<td></td>
<td></td>
<td>DRO-One</td>
<td>45.34</td>
<td>64.41</td>
<td>70.14</td>
<td>66.53</td>
<td>43.59</td>
<td>63.42</td>
<td>69.33</td>
</tr>
<tr>
<td></td>
<td></td>
<td>DRO-Random</td>
<td>45.73</td>
<td>64.66</td>
<td>71.00</td>
<td>64.16</td>
<td>42.40</td>
<td>64.38</td>
<td>70.74</td>
</tr>
<tr>
<td></td>
<td></td>
<td>DRO-Language</td>
<td>44.73</td>
<td>65.16</td>
<td>71.08</td>
<td>57.54</td>
<td>40.57</td>
<td>62.68</td>
<td>69.78</td>
</tr>
<tr>
<td></td>
<td>mBERT</td>
<td>ERM</td>
<td>56.81</td>
<td>77.03</td>
<td>81.65</td>
<td>34.49</td>
<td>46.87</td>
<td>72.55</td>
<td>79.49</td>
</tr>
<tr>
<td></td>
<td></td>
<td>DRO-One</td>
<td>47.49</td>
<td>65.84</td>
<td>70.97</td>
<td>76.57</td>
<td>36.17</td>
<td>56.74</td>
<td>64.51</td>
</tr>
<tr>
<td></td>
<td></td>
<td>DRO-Random</td>
<td>47.85</td>
<td>65.88</td>
<td>70.93</td>
<td>74.87</td>
<td>37.66</td>
<td>58.37</td>
<td>65.41</td>
</tr>
<tr>
<td></td>
<td></td>
<td>DRO-Language</td>
<td>45.77</td>
<td>64.77</td>
<td>70.39</td>
<td>65.55</td>
<td>33.94</td>
<td>55.01</td>
<td>62.09</td>
</tr>
<tr>
<td>Naija</td>
<td>BERT</td>
<td>ERM</td>
<td>73.72</td>
<td>88.62</td>
<td>91.99</td>
<td>28.14</td>
<td>71.38</td>
<td>87.33</td>
<td>90.94</td>
</tr>
<tr>
<td></td>
<td></td>
<td>DRO-One</td>
<td>64.28</td>
<td>79.37</td>
<td>83.95</td>
<td>61.81</td>
<td>59.86</td>
<td>77.00</td>
<td>81.60</td>
</tr>
<tr>
<td></td>
<td></td>
<td>DRO-Random</td>
<td>63.72</td>
<td>79.57</td>
<td>83.92</td>
<td>60.31</td>
<td>59.31</td>
<td>75.55</td>
<td>80.29</td>
</tr>
<tr>
<td></td>
<td></td>
<td>DRO-Language</td>
<td>63.58</td>
<td>79.48</td>
<td>84.29</td>
<td>56.83</td>
<td>59.74</td>
<td>77.08</td>
<td>81.84</td>
</tr>
<tr>
<td>Haitian</td>
<td>BERT</td>
<td>ERM</td>
<td>72.96</td>
<td>87.58</td>
<td>91.15</td>
<td>31.77</td>
<td>70.42</td>
<td>87.58</td>
<td>91.42</td>
</tr>
<tr>
<td></td>
<td></td>
<td>DRO-One</td>
<td>63.72</td>
<td>78.36</td>
<td>82.77</td>
<td>76.24</td>
<td>60.78</td>
<td>77.07</td>
<td>81.78</td>
</tr>
<tr>
<td></td>
<td></td>
<td>DRO-Random</td>
<td>63.52</td>
<td>77.77</td>
<td>82.18</td>
<td>74.53</td>
<td>61.02</td>
<td>78.01</td>
<td>82.88</td>
</tr>
<tr>
<td></td>
<td></td>
<td>DRO-Language</td>
<td>63.13</td>
<td>78.16</td>
<td>82.80</td>
<td>71.77</td>
<td>60.73</td>
<td>77.37</td>
<td>82.25</td>
</tr>
<tr>
<td></td>
<td>mBERT</td>
<td>ERM</td>
<td>66.06</td>
<td>80.45</td>
<td>84.30</td>
<td>69.25</td>
<td>55.58</td>
<td>72.49</td>
<td>78.72</td>
</tr>
<tr>
<td></td>
<td></td>
<td>DRO-One</td>
<td>50.35</td>
<td>65.05</td>
<td>69.10</td>
<td>174.45</td>
<td>35.86</td>
<td>51.60</td>
<td>56.72</td>
</tr>
<tr>
<td></td>
<td></td>
<td>DRO-Random</td>
<td>48.63</td>
<td>64.03</td>
<td>67.78</td>
<td>172.26</td>
<td>32.54</td>
<td>48.31</td>
<td>53.60</td>
</tr>
<tr>
<td></td>
<td></td>
<td>DRO-Language</td>
<td>49.14</td>
<td>64.24</td>
<td>68.69</td>
<td>167.90</td>
<td>34.59</td>
<td>49.34</td>
<td>55.45</td>
</tr>
</tbody>
</table>

### Table 10: Full results for pretrained baselines.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Model</th>
<th>Strategy</th>
<th>P@1</th>
<th>P@5</th>
<th>P@10</th>
<th>PLL</th>
<th>P_D@1</th>
<th>P_D@5</th>
<th>P_D@10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Singlish</td>
<td>BERT</td>
<td>ERM</td>
<td>23.94</td>
<td>38.49</td>
<td>45.09</td>
<td>76.01</td>
<td>21.09</td>
<td>36.65</td>
<td>42.22</td>
</tr>
<tr>
<td></td>
<td></td>
<td>DRO-One</td>
<td>14.30</td>
<td>23.12</td>
<td>27.03</td>
<td>92.97</td>
<td>10.23</td>
<td>23.57</td>
<td>29.85</td>
</tr>
<tr>
<td>Nigerian Pidgin</td>
<td>BERT</td>
<td>ERM</td>
<td>22.79</td>
<td>34.04</td>
<td>39.88</td>
<td>142.66</td>
<td>10.92</td>
<td>18.07</td>
<td>22.96</td>
</tr>
<tr>
<td></td>
<td></td>
<td>DRO-One</td>
<td>14.90</td>
<td>26.34</td>
<td>31.87</td>
<td>153.54</td>
<td>8.08</td>
<td>16.24</td>
<td>20.72</td>
</tr>
<tr>
<td>Haitian Creole</td>
<td>BERT</td>
<td>ERM</td>
<td>18.84</td>
<td>30.60</td>
<td>37.59</td>
<td>177.40</td>
<td>5.65</td>
<td>11.89</td>
<td>16.29</td>
</tr>
<tr>
<td></td>
<td></td>
<td>DRO-One</td>
<td>11.96</td>
<td>22.39</td>
<td>27.96</td>
<td>175.14</td>
<td>7.10</td>
<td>12.20</td>
<td>16.76</td>
</tr>
</tbody>
</table>

Table 9: Full results for Creole-Only experiments.

Table 10: Full results for pretrained baselines.