On Language Models for Creoles

Heather Lent[c] **Emanuele Bugliarello**[c] **Miryam de Lhoneux**[c,υ,ð]
Chen Qiu[ɯ] **Anders Søgaard**[c]
[c] University of Copenhagen, Denmark
[υ] Uppsala University, Sweden [ð] KU Leuven, Belgium
[ɯ] 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*,[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
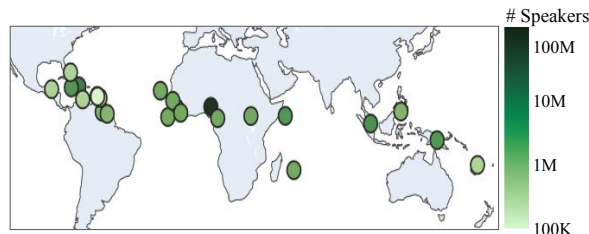

Figure 1: Creoles with a minimum of a hundred thousand speakers are shown here (Hawaiian Pidgin not pictured). Approximately 180 million creole speakers are represented in this map. Data extracted from https://en.wikipedia.org/wiki/List_of_creole_languages.

hundred million speakers of creole languages world wide (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

---

[1] A pidgin is a grammatically simplified language that develops between two or more groups that do not have a language in common. Both pidgins and creoles are sometimes referred to as *contact* languages.

Tamil		Mandarin(我们)	Cantonese(拍拖)	English	Malay	Eng	Malay	Hokkien/ Hakka(店)	X
Dey	,	wǒ men	paktor	always	makan	at	kopi	tiam	one
Hey	,	we	date	always	eat	at	coffee	shop	<INTJ>

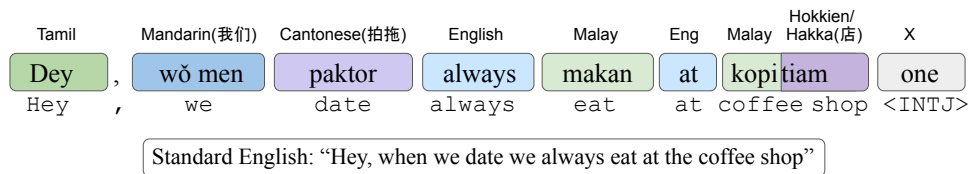Standard English: "Hey, when we date we always eat at the coffee shop"

Figure 2: Example sentence in Singlish featuring multilingual vocabulary, Chinese-style topic prominence combined with a subordinate clause with English word order, and a final interjection representing a discourse particle; a common feature of Singlish. Example from https://languagelog.ldc.upenn.edu/nll/?p=25758.

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,[2] 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 straight-forward 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 Murawaki (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 "restructurer", 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.

---

[2]Creole languages clearly differ though in the dynamics that affect their drift. For example, Yakpo (2021) discuss two seemingly similar creole languages, Krio (Sierra Leone) and Pichi (Equatorial Guinea). Both creoles have English as their lexifier, but while Krio is spoken alongside English, Pichi is spoken alongside Spanish. The two creoles, as a consequence, exhibit a clear difference. Krio has converged increasingly toward English, while Pichi has neither converged toward English nor Spanish.

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 from 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 transla-

Language	Source	Domain
en, fr, es, pt, yo, zh, ta	WMT-News 2020	news
ms	Malay 30k News	news
Nigerian Pidgin	PidginUNMT Corpus	news
Singlish	Singapore SMS Corpus	sms
Haitian Creole	Disaster Response Corpus	sms

Table 1: Data resources utilized in our experiments.

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

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.

tion 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-jussà and Banchs, 2011) as well as spell checking and data cleaning (Stymne, 2011).

4   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.

Data splits We begin with the creole datasets noted in Tab. 1, and combine them with data of other higher-resource languages that have been influential to the creole. We combine a fixed number of these examples into a MIXED-LANGUAGE dataset, as described in Tab. 2. The MIXED-LANGUAGE dataset for each creole includes in-

---

[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.[4] 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
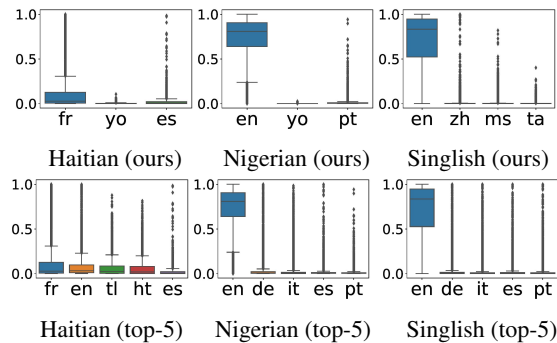
Figure 3: Distributions of identified languages across the CREOLE-ONLY test set. **Top:** distributions for the influential languages included in MIXED-LANGUAGE. **Bottom:** distributions of the five languages that had the highest prediction scores for each creole, where we see a bias towards European languages.

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.

5 Experiments

In this section, we detail our experimental setups. We make our code and models publicly available.[5]

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$_{\text{Base}}$ 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 of 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.[6] 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 $\log \mathbb{P}_{\text{MLM}}(w_t | \mathbf{w}_{\backslash t})$ of each token w_t in a sentence $\mathbf{w} = \langle w_1, \ldots, w_T \rangle$. These are obtained in BERT

---

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

[6]Nigerian Pidgin: http://naijalingo.com/. Singlish: http://www.mysmu.edu/faculty/jacklee/. Haitian Creole: https://kreyol.com/dictionary.html.

		Nigerian Pidgin			Singlish			Haitian Creole		
BERT		P@1	P$_D$@1	PLL	P@1	P$_D$@1	PLL	P@1	P$_D$@1	PLL
	Pretrained	22.79	10.92	142.65	23.94	21.09	76.01	18.84	5.65	177.40
MIXED	ERM	**63.83**	**59.97**	**42.41**	**46.77**	**42.89**	**41.06**	**68.09**	**43.35**	**55.04**
	DRO-One	60.99	56.76	52.51	44.23	40.73	49.18	57.04	36.73	121.51
	DRO-Random	60.40	56.33	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	57.55	36.69	118.85
C-ONLY	ERM	**73.72**	**71.38**	**28.14**	**53.80**	**51.26**	**34.22**	**73.15**	**55.50**	**55.51**
	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

Table 3: Intrinsic evaluation: Precision@1 (P@1), Precision@1 for words in our creole dictionary (P$_D$@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.

by replacing w_t with the special `[MASK]` token. Here, we report the mean score given by:

$$\text{PLL} = \frac{1}{|\mathcal{C}|} \sum_{\mathbf{w} \in \mathcal{C}} \frac{1}{|\mathbf{w}|} \sum_{w_t \in \mathbf{w}} \log \mathbb{P}_{\text{MLM}}(w_t | \mathbf{w}_{\backslash t}; \theta), \quad (1)$$

where $\mathcal{C}$ denotes the evaluation corpus, and θ 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 checkpoints 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

Intrinsic evaluation The main finding of the intrinsic evaluation is that ERM outperforms DRO

		Nigerian Pidgin		Singlish
BERT		NER [F$_1$]	UPOS [Acc]	UPOS [Acc]
MIXED	ERM	87.86	98.00	**91.24**
	DRO-Language	**88.40**	**98.06**	90.22
C-ONLY	ERM	**87.98**	**98.04**	**91.17**
	DRO-Language	87.12	97.98	90.44

Table 4: Extrinsic evaluation. Similar performance on downstream tasks across all models demonstrate show that language model training did *not* benefit significantly from neither DRO nor data in related languages.

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 patters 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 conclusioins from.

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

		Nigerian Pidgin		
BERT	**Size**	**P@1**	**P$_D$@1**	**PLL**
ERM	Tiny	31.31	26.12	110.23
	Small	47.39	46.75	77.47
	Base	63.83	59.97	42.41
DRO-Language	Tiny	31.00	23.09	99.70
	Small	43.00	37.75	82.50
	Base	60.40	54.80	54.17

Table 5: Over-parameterization experiments with MIXED-LANGUAGE Nigerian Pidgin English data. Smaller sized models do not benefit DRO over ERM.

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 BERT$_{Tiny}$, BERT$_{Small}$ (Jiao et al., 2020), and BERT$_{Base}$. 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 BERT$_{Small}$ 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

		Nigerian Pidgin		
BERT	**Weight Decay**	**P@1**	**P$_D$@1**	**PLL**
ERM	0.01	47.39	46.75	77.47
DRO-Language	0.01	43.00	37.75	82.50
	0.05	42.86	38.47	83.03
	0.10	43.00	38.74	81.80
	0.30	42.70	39.53	81.94

Table 6: Regularization experiments on MIXED-LANGUAGE Nigerian Pidgin data, based on BERT$_{Small}$.

Language	**Domain-1**	**Domain-2**	**PAD**
English	Disaster Response Corpus	Newswire	1.75
Haitian Creole	Disaster Response Corpus	Newswire	1.47
English	EWT-UD	NUD	1.04
Nigerian	UNMT	NUD	1.28

Table 7: Proxy $\mathcal{A}$-distance (PAD) scores on parallel (Haitian) or near-parallel (Nigerian) data. PAD is proportional to domain classification error; hence, large distances mean high domain divergence. Our results suggest that creole languages do *not* exhibit significantly more drift than other languages.

investigate the matter of creole stability, we follow (Ben-David et al., 2007) and calculate the proxy $\mathcal{A}$-distance (PAD) between different domains of creole data (see Tab. 7). Specifically, we train an SVM on the BERT encodings.[7] Our $\mathcal{A}$-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 minimiziation is superior. We show that this is not the result of over-parameterization or lack of regularization, but in-

---

[7]Our code is adapted from https://github.com/rpryzant/proxy-a-distance.

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

References

Enoch Oladé Aboh. 2015. *The emergence of hybrid grammars : language contact and change / Enoch Oladé Aboh.* Cambridge approaches to language contact. Cambridge University Press, Hastings, England.

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 Muhammad, Chris Chinenye Emezue, Joyce Nakatumba-Nabende, Perez Ogayo, Anuoluwapo Aremu, Catherine Gitau, Derguene Mbaye, Jesujoba Alabi, Seid Muhie Yimam, Tajuddeen 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.

Željko Agić and Ivan Vulić. 2019. JW300: A wide-coverage parallel corpus for low-resource languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3204–3210, Florence, Italy. Association for Computational Linguistics.

Orevaoghene Ahia and Kelechi Ogueji. 2020. Towards supervised and unsupervised neural machine translation baselines for nigerian pidgin. *ArXiv*, abs/2003.12660.

Daniel Ajisafe, O. Adegboro, Esther Oduntan, and T. Arulogun. 2020. Towards end-to-end training of automatic speech recognition for nigerian pidgin. *ArXiv*, abs/2010.11123.

Rajiv Bajpai, Soujanya Poria, Danyuan Ho, and Erik Cambria. 2017. Developing a concept-level knowledge base for sentiment analysis in singlish. *CoRR*, abs/1707.04408.

Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. 2007. Analysis of representations for domain adaptation. In *Advances in Neural Information Processing Systems*, volume 19. MIT Press.

A. Ben-Tal, D. D. Hertog, A. D. Waegenaere, B. Melenberg, and G. Rennen. 2013. Robust solutions of optimization problems affected by uncertain probabilities. *Manag. Sci.*, 59:341–357.

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

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Omar F. Zaidan, editors. 2011. *Proceedings of the Sixth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Edinburgh, Scotland.

Bernard Caron, Marine Courtin, Kim Gerdes, and Sylvain Kahane. 2019. A surface-syntactic UD treebank for Naija. In *Proceedings of the 18th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2019)*, pages 13–24, Paris, France. Association for Computational Linguistics.

Özlem Çetinoğlu, Sarah Schulz, and Ngoc Thang Vu. 2016. Challenges of computational processing of code-switching. In *Proceedings of the Second Workshop on Computational Approaches to Code Switching*, pages 1–11, Austin, Texas. Association for Computational Linguistics.

Ethan C. Chau, Lucy H. Lin, and Noah A. Smith. 2020. Parsing with multilingual BERT, a small corpus, and a small treebank. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1324–1334, Online. Association for Computational Linguistics.

T. Chen and Kan Min-Yen. 2015. The national university of singapore sms corpus.

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.

A. Seza Doğruöz, Sunayana Sitaram, Barbara E. Bullock, and Almeida Jacqueline Toribio. 2021. A survey of code-switching: Linguistic and social perspectives for language technologies. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1654–1666, Online. Association for Computational Linguistics.

Geli Fei and Bing Liu. 2016. Breaking the closed world assumption in text classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 506–514, San Diego, California. Association for Computational Linguistics.

Vitaly Feldman and Chiyuan Zhang. 2020. What neural networks memorize and why: Discovering the long tail via influence estimation. In *Advances in Neural Information Processing Systems*, volume 33, pages 2881–2891. Curran Associates, Inc.

Gregory Furman and G. Nitschke. 2020. Evolving an artificial creole. *Proceedings of the 2020 Genetic and Evolutionary Computation Conference Companion*.

Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. 2018. Fairness without demographics in repeated loss minimization. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1929–1938. PMLR.

Sanjika Hewavitharana, Nguyen Bach, Qin Gao, Vamshi Ambati, and Stephan Vogel. 2011. CMU Haitian Creole-English translation system for WMT 2011. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 386–392, Edinburgh, Scotland. Association for Computational Linguistics.

Danyuan Ho, Diyana Hamzah, Soujanya Poria, and Erik Cambria. 2018. Singlish senticnet: A concept-based sentiment resource for singapore english. In *2018 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 1285–1291.

John Holm. 2000. *Social factors*, Cambridge Textbooks in Linguistics, page 68–105. Cambridge University Press.

Chang Hu, P. Resnik, Y. Kronrod, Vladimir Eidelman, Olivia Buzek, and B. Bederson. 2011a. The value of monolingual crowdsourcing in a real-world translation scenario: Simulation using haitian creole emergency sms messages. In *WMT@EMNLP*.

Chang Hu, Philip Resnik, Yakov Kronrod, Vladimir Eidelman, Olivia Buzek, and Benjamin B. Bederson. 2011b. The value of monolingual crowdsourcing in a real-world translation scenario: Simulation using haitian creole emergency sms messages. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 399–404, Edinburgh, Scotland. Association for Computational Linguistics.

Fredrik Jansson, Mikael Parkvall, and Pontus Strimling. 2015. Modeling the evolution of creoles. *Language Dynamics and Change*, 5(1):1 – 51.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2020. TinyBERT: Distilling BERT for natural language understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4163–4174, Online. Association for Computational Linguistics.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Sara Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. 2020. WILDS: A benchmark of in-the-wild distribution shifts. *CoRR*, abs/2012.07421.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton A. Earnshaw, Imran S. Haque, Sara Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. 2021. Wilds: A benchmark of in-the-wild distribution shifts.

Sibylle Kriegel. 2016. 2014. pidgins and creoles beyond africa-europe encounters, written by isabelle buchstaller, anders holmberg and mohammed al-moaily. *Journal of Language Contact*, 9(3):576 – 579.

Siaw Ling Lo, Erik Cambria, Raymond Chiong, and David Cornforth. 2016. A multilingual semi-supervised approach in deriving singlish sentic patterns for polarity detection. *Knowledge-Based Systems*, 105:236–247.

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

Robert Munro. 2010. Crowdsourced translation for emergency response in haiti: the global collaboration of local knowledge. In *In Relief 2.0 in Haiti*.

Robert Munro. 2013. Crowdsourcing and the crisis-affected community - lessons learned and looking forward from mission 4636. *Inf. Retr.*, 16(2):210–266.

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

Makoto Nakamura, Takashi Hashimoto, and Satoshi Tojo. 2009. Prediction of creole emergence in spatial language dynamics. In *Language and Automata Theory and Applications*, pages 614–625, Berlin, Heidelberg. Springer Berlin Heidelberg.

Innocent Ndubuisi-Obi, S. Ghosh, and David Jurgens. 2019. Wetin dey with these comments? modeling sociolinguistic factors affecting code-switching behavior in nigerian online discussions. In *ACL*.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *LREC*, pages 4034–4043, Marseille, France. European Language Resources Association.

Kelechi Ogueji and Orevaoghene Ahia. 2019. Pidginunmt: Unsupervised neural machine translation from west african pidgin to english. *ArXiv*, abs/1912.03444.

Yonatan Oren, Shiori Sagawa, Tatsunori B. Hashimoto, and Percy Liang. 2019. Distributionally robust language modeling. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4227–4237, Hong Kong, China. Association for Computational Linguistics.

Wuraola Fisayo Oyewusi, Olubayo Adekanmbi, and O. Akinsande. 2020. Semantic enrichment of nigerian pidgin english for contextual sentiment classification. *ArXiv*, abs/2003.12450.

Wuraola Fisayo Oyewusi, Olubayo Adekanmbi, Ifeoma Okoh, Vitus Onuigwe, Mary Idera Salami, Opeyemi Osakuade, Sharon Ibejih, and Usman Abdullahi Musa. 2021. Naijaner : Comprehensive named entity recognition for 5 nigerian languages.

Mikael Parkvall. 2008. The simplicity of creoles in a cross-linguistic perspective.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. *CoRR*, abs/1912.01703.

Peter Patrick. 1999. *Urban Jamaican Creole. Variation in the Mesolect*. Number G17 in Varieties of English Around the World. John Benjamins Publishing Co., Amsterdam & Phildadelphia.

I. Plag. 2009. Creoles as interlanguages : Phonology. *Journal of Pidgin and Creole Languages*, 24:119–138.

Adithya Pratapa, Gayatri Bhat, Monojit Choudhury, Sunayana Sitaram, Sandipan Dandapat, and Kalika Bali. 2018. Language modeling for code-mixing: The role of linguistic theory based synthetic data. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1543–1553, Melbourne, Australia. Association for Computational Linguistics.

Marta R. Costa-jussà and Rafael E. Banchs. 2011. The bm-i2r haitian-créole-to-english translation system description for the wmt 2011 evaluation campaign. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 452–456, Edinburgh, Scotland. Association for Computational Linguistics.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. 2019. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *CoRR*, abs/1911.08731.

Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. 2020. Masked language model scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2699–2712, Online. Association for Computational Linguistics.

Walter J. Scheirer, Anderson de Rezende Rocha, Archana Sapkota, and Terrance E. Boult. 2013. Toward open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(7):1757–1772.

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

Payal Kushal Shah-Sanghavi. 2017. Should creoles be made official languages and / or media of instruction in countries where they are the first language of the majority of the population? *IOSR Journal Of Humanities And Social Science*, 22:19–25.

Joonbo Shin, Yoonhyung Lee, and Kyomin Jung. 2019. Effective sentence scoring method using bert for speech recognition. In *Proceedings of The Eleventh*

Asian Conference on Machine Learning, volume 101 of *Proceedings of Machine Learning Research*, pages 1081–1093, Nagoya, Japan. PMLR.

Sara Stymne. 2011. Spell checking techniques for replacement of unknown words and data cleaning for haitian creole sms translation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 470–477, Edinburgh, Scotland. Association for Computational Linguistics.

Samson Tan, Shafiq Joty, Lav Varshney, and Min-Yen Kan. 2020. Mind your inflections! Improving NLP for non-standard Englishes with Base-Inflection Encoding. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5647–5663, Online. Association for Computational Linguistics.

Titi Ufomata. 1999. Major and minor languages in complex linguistic ecologies: the nigerian experience. *International Journal of Educational Development*, 19(4):315–322.

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. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 176–197, Online. Association for Computational Linguistics.

Alex Wang and Kyunghyun Cho. 2019. Bert has a mouth, and it must speak: Bert as a markov random field language model. *arXiv preprint arXiv:1902.04094*.

Hongmin Wang, Jie Yang, and Yue Zhang. 2019. From genesis to creole language: Transfer learning for singlish universal dependencies parsing and pos tagging. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 19(1).

Hongmin Wang, Yue Zhang, GuangYong Leonard Chan, Jie Yang, and Hai Leong Chieu. 2017. Universal Dependencies parsing for colloquial singaporean English. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1732–1744, Vancouver, Canada. Association for Computational Linguistics.

Donald Winford. 1999. Variation theory: a view from creole continua.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *CoRR*, abs/1910.03771.

Kofi Yakpo. 2021. Two types of language contact involving english creoles: Why krio (sierra leone) has evolved more towards english than its relative pichi (equatorial guinea) towards spanish. *English Today*, page 1–12.

Yinhe Zheng, Guanyi Chen, and Minlie Huang. 2020. Out-of-domain detection for natural language understanding in dialog systems.

A Appendix

Singlish							
Model	Strategy	P@1	P@5	P@10	PLL	$P_D@1$	$P_D@5$	$P_D@10$
BERT	ERM	46.77	68.89	74.34	41.07	42.89	66.76	74.17
	DRO-One	44.23	64.73	71.90	49.18	40.73	63.05	70.13
	DRO-Random	43.33	65.63	71.58	49.14	39.07	61.02	68.42
	DRO-Language	43.19	64.80	71.22	48.88	39.57	61.54	70.34
mBERT	ERM	47.78	68.82	75.77	42.26	37.00	61.71	70.85
	DRO-One	44.37	65.13	72.54	50.99	33.71	57.79	65.72
	DRO-Random	44.34	65.95	72.44	50.39	35.25	59.20	66.93
	DRO-Language	43.69	64.91	71.61	50.49	33.45	58.91	67.42
Naija							
Model	Strategy	P@1	P@5	P@10	PLL	$P_D@1$	$P_D@5$	$P_D@10$
BERT	ERM	63.83	80.52	85.44	42.41	59.97	78.72	83.93
	DRO-One	60.99	77.52	82.94	52.51	56.76	76.21	81.94
	DRO-Random	60.40	78.44	82.88	52.69	56.33	75.27	81.17
	DRO-Language	60.40	77.40	82.69	54.18	54.80	74.48	80.32
mBERT	ERM	62.68	80.52	85.55	44.98	62.19	82.25	87.08
	DRO-One	60.76	77.74	82.88	58.15	56.43	77.61	82.67
	DRO-Random	60.34	77.23	82.24	57.90	56.70	77.50	82.51
	DRO-Language	58.88	76.56	81.73	59.91	54.99	76.63	82.50
Haitian							
Model	Strategy	P@1	P@5	P@10	PLL	$P_D@1$	$P_D@5$	$P_D@10$
BERT	ERM	68.09	82.98	87.34	55.05	43.35	63.89	71.35
	DRO-One	57.04	71.12	75.58	121.51	36.73	52.55	58.25
	DRO-Random	57.65	71.53	75.79	119.17	36.16	50.63	56.38
	DRO-Language	57.55	71.23	75.28	118.85	36.69	50.48	55.89
mBERT	ERM	60.79	76.70	81.56	60.27	46.35	64.56	70.96
	DRO-One	51.06	65.45	69.71	148.12	34.57	49.30	55.61
	DRO-Random	50.86	65.05	69.50	146.18	34.52	49.58	55.00
	DRO-Language	50.15	64.54	69.40	145.97	33.55	48.21	55.08

Table 8: Full results for Mixed-Language experiments.

Singlish

Model	Strategy	P@1	P@5	P@10	PLL	P_D@1	P_D@5	P_D@10
BERT	ERM	53.80	75.02	80.36	34.22	51.26	74.09	80.15
	DRO-One	45.34	64.41	70.14	66.53	43.59	63.42	69.33
	DRO-Random	45.73	64.66	71.00	64.16	42.40	64.38	70.74
	DRO-Language	44.73	65.16	71.08	57.54	40.57	62.68	69.78
mBERT	ERM	56.81	77.03	81.65	34.49	46.87	72.55	79.49
	DRO-One	47.49	65.84	70.97	76.57	36.17	56.74	64.51
	DRO-Random	47.85	65.88	70.93	74.87	37.66	58.37	65.41
	DRO-Language	45.77	64.77	70.39	68.55	33.94	55.01	62.09

Naija

Model	Strategy	P@1	P@5	P@10	PLL	P_D@1	P_D@5	P_D@10
BERT	ERM	73.72	88.62	91.99	28.14	71.38	87.33	90.94
	DRO-One	64.28	79.37	83.95	61.81	59.86	77.00	81.60
	DRO-Random	63.72	79.57	83.92	60.31	59.31	75.55	80.29
	DRO-Language	63.58	79.48	84.29	56.83	59.74	77.08	81.84
mBERT	ERM	72.96	87.58	91.15	31.77	70.42	87.58	91.42
	DRO-One	63.72	78.36	82.77	76.24	60.78	77.07	81.78
	DRO-Random	63.52	77.77	82.18	74.53	61.02	78.01	82.88
	DRO-Language	63.13	78.16	82.80	71.77	60.73	77.37	82.25

Haitian

Model	Strategy	P@1	P@5	P@10	PLL	P_D@1	P_D@5	P_D@10
BERT	ERM	73.15	86.12	88.55	55.51	55.50	71.76	77.94
	DRO-One	58.16	71.23	75.48	144.47	36.91	51.39	56.04
	DRO-Random	57.65	70.52	75.38	142.04	37.41	52.55	57.72
	DRO-Language	56.94	71.33	74.97	138.60	35.50	49.66	55.03
mBERT	ERM	66.06	80.45	84.30	69.25	55.58	72.49	78.72
	DRO-One	50.35	65.05	69.10	174.45	35.86	51.60	56.72
	DRO-Random	48.63	64.03	67.78	172.26	32.54	48.31	53.60
	DRO-Language	49.14	64.24	68.69	167.90	34.59	49.34	55.45

Table 9: Full results for Creole-Only experiments.

Dataset	Model	P@1	P@5	P@10	PLL	P_D@1	P_D@5	P_D@10
Singlish	BERT	23.94	38.49	45.09	76.01	21.09	36.65	42.22
	mBERT	14.30	23.12	27.03	92.97	10.23	23.57	29.85
Nigerian Pidgin	BERT	22.79	34.04	39.88	142.66	10.92	18.07	22.96
	mBERT	14.90	26.34	31.87	153.54	8.08	16.24	20.72
Haitian Creole	BERT	18.84	30.60	37.59	177.40	5.65	11.89	16.29
	mBERT	11.96	22.39	27.96	175.14	7.10	12.20	16.76

Table 10: Full results for pretrained baselines.