What a Creole Wants, What a Creole Needs

Heather Lent[1], Kelechi Ogueji[2], Miryam de Lhoneux[1,3,4], Orevaoghene Ahia[5], Anders Søgaard[1]

[1] University of Copenhagen, Denmark
[2] University of Waterloo, Canada
[3] Uppsala University, Sweden
[4] KU Lueven, Belgium
[5] University of Washington, United States
{hcl, ml, soegaard}@di.ku.dk
kelechi.ogueji@uwaterloo.ca, oahia@cs.washington.edu

Abstract

In recent years, the natural language processing (NLP) community has given increased attention to the disparity of efforts directed towards high-resource languages over low-resource ones. Efforts to remedy this delta often begin with translations of existing English datasets into other languages. However, this approach ignores that different language communities have different needs. We consider a group of low-resource languages, *Creole* languages. Creoles are both largely absent from the NLP literature, and also often ignored by society at large due to stigma, despite these languages having sizable and vibrant communities. We demonstrate, through conversations with Creole experts and surveys of Creole-speaking communities, how the things needed from language technology can change dramatically from one language to another, even when the languages are considered to be very similar to each other, as with Creoles. We discuss the prominent themes arising from these conversations, and ultimately demonstrate that useful language technology cannot be built without involving the relevant community.

Keywords: natural language processing, low-resource languages, Creole

1. Introduction

The field of natural language processing (NLP) has become aware that most of the world's languages are unfortunately under-represented, or entirely absent, from the field's body of work (Joshi et al., 2020). In recent years, there has been a push in efforts to ameliorate this discrepancy (Nekoto et al., 2020; Mirzakhalov et al., 2021; Ogueji et al., 2021). Among these low-resourced languages[1] are *Creole* languages, which are particularly under-resourced due to barriers like societal stigma (Siegel, 1999), despite the fact that these languages are spoken by many people globally. One line of work has focused on creating datasets for low-resource languages via the translation of existing high-resource language datasets (Conneau et al., 2018; Artetxe et al., 2020b; Budur et al., 2020). Despite the popularity of this method, it poses several issues, which can negatively affect the communities of these low-resource languages. One such issue lies in translation artifacts, which have been shown to have notable impacts on the performance of models trained with such datasets (Artetxe et al., 2020a). Furthermore, translated datasets are often simplified and unnatural, a phenomenon referred to as *translationese* (Volansky et al., 2013). This has also been shown to adversely affect the evaluation of machine translation models (Graham et al., 2020). Creoles, too, are not immune to the shortcomings of this approach. Moreover, many trans-

lated datasets will simply not be relevant to communities speaking a Creole language, as concepts relevant to the original high-resource source language are subsequently translated into the low-resource language, despite being irrelevant to people or cultures speaking the language (Liu et al., 2021). For example, sentences about American football or the American Thanksgiving holiday are simply not relevant or necessary for speakers of Creole languages. The same mismatch also applies to other more geographical-specific domain information present in the data, such as landmarks or landscapes. All of these show that, while there may be good intentions behind this approach, it could potentially lead to poor models for speakers of low-resource languages and even to the creation of tools of little use or relevance for Creole speakers.

Meanwhile, works such as Hu et al. (2011) concretely demonstrate how crowd-sourcing data from target-language speakers, even if monolingual, leads to improved results for statistical machine translation systems. While these findings are not up to date with contemporary neural machine translation, involving native speakers minimizes the risk of having non-relevant examples included in the dataset. However, as the authors also note, there can be considerable logistical difficulties of finding native speakers to contribute, even when offering payment. And even if one manages to recruit paid speakers, a large problem still remains: the underlying exploitative nature of treating language speaking communities like data resources to be mined. Bird (2020) discusses in detail these foundational problems within the language technology community, and

---

[1]This term is often largely ambiguous, and all "low-resource" languages should not be conflated together into one large group, but rather considered independently, in the context of its speakers, their culture, and their needs.

how, in order to break the cycle of harmful colonialism in our science, we must fundamentally change the relationship between researchers and the language-speaking communities. But the only way we can learn this, claims Bird (2020), is by establishing a respectful, "feedback/collaboration loop", and necessarily involving community members in our research.

Following the work of Bird (2020), in this work, we focus on the problem of creating resources for low-resource languages, in this case Creoles, and the inherent presupposition by researchers of what technologies are indeed wanted and needed by the communities speaking those languages. While many researchers may assume that the "best-case scenario" for all languages would be to have all language technologies *equally* available, the fact of the matter is that many communities have very specific wants and needs of language technology, as well as language technologies that are notably unwelcome, even though they are a commonplace for high-resource languages. Disregarding the needs of a language community can lead to misuse of finite resources on creating unnecessary datasets or technologies while leaving the community's highest priorities neglected. And finally, when researchers assume what technologies are wanted on the behalf of a community, it inherently alienates that community and takes away their agency (Bird, 2020). In this work, we explore how the needs of different Creole-speaking communities vary wildly from one another, and we demonstrate the need to establish respectful relationships with experts and communities, in order to make truly useful language technology.

Our contributions in this work are as follows:

- We present a survey of Creole NLP, and discussion of features from Creole languages that present unique challenges to existing NLP workflows.

- We discuss important considerations, gleaned from conversations with experts, and a survey of Creole language speakers.

- We propose a Creole continuum for language technology, as a guiding framework of research considerations, to help NLP researchers planning to work on Creoles.

2. Background

Today, Creole languages are spoken widely throughout the Caribbean and West Africa, as well as parts of South America, Asia, Australia and the Pacific. Creoles have long captured the attention of linguists due to their unique, and sometimes tragic[2], histories with regards to language evolution. Typically, Creole languages originate from situations in which multiple different languages have come into close contact with each other

(Thomason and Kaufman, 1992). The exact process of how a Creole language is "born" (i.e. Creole genesis), as well as discussion of which linguistic features a Creole inherited from the various "parent" languages, have been the subject of intense and ongoing linguistic debate for decades (Alleyne, 1971; Bickerton, 1984; Muysken and Smith, 1986; Sessarego, 2020). On one hand, some believe that Creoles themselves form a unique typological class of languages, with a separate place on the phylogenetic tree of languages (i.e. Creole exceptionalism, Bickerton (1984)). Linguists supporting Creole exceptionalism typically claim that Creoles are more simple than other languages (Parkvall and others, 2008), for example, lacking in complex morphology (McWhorter, 1998). On the other hand, others argue that there are no grounds to claim that Creole evolution is especially different from the language evolution of so-called "normal" (DeGraff, 2003; DeGraff, 2005). And indeed, Creoles do exhibit behaviors just as complex as non-Creole languages (DeGraff, 2001), including complex morphology (Henri et al., 2020).

Moreover, some criticisms of Creole exceptionalism also examine how the history of Creole studies itself has unfortunately been riddled with discrimination and racism (DeGraff, 2005). In the past, Creoles were often considered to be something short of a full-fledged language (or, more harshly, "degenerate variants or dialects of their parent languages"[3]). According to Kouwenberg and Singler (2009), "A part of the legacy of slavery in the Caribbean and elsewhere has been the stigmatization of the languages associated with slaves ... [the] willingness to apply the concept of linguistic relativism – whereby every language is understood to be complete and valid – may have been extended to Hopi and Hausa[4], but it generally stopped short of being extended to Creoles." In line with this, in this work, we hope to raise awareness in the NLP community about why Creoles are important to work with. Beyond being the subject of vibrant linguistic debate, Creoles are often ignored when it comes to language technology, which puts speakers of already often stigmatized languages at a further disadvantage. For the remainder of this section, we will present a survey of existing Creole datasets, a summary of works published on NLP for Creoles, and finally end this section with a discussion of some specific features of Creole languages that are notable within the context of NLP.

2.1. Creole Data and Creole NLP

In this section we will detail existing resources and datasets for Creole languages (including those which are now seemingly defunct), as well as discuss related works actively focused on NLP for Creoles.

---

[2]For example, Caribbean Creoles resulted from the displacement of African peoples in the Atlantic slave trade.

[3]https://en.wikipedia.org/wiki/Creole_language#Overview

[4]Hopi is an Native American indigenous language from Arizona, United States; Hausa is a Chadic language, spoken in West and Central Africa.

Verified Resources Although Creole languages are in general very low-resourced, the datasets that do exist vary widely from task to task, as well as from language to language. Hagemeijer et al. (2014) presents an extensive overview of Creole data resources through 2014 for a wide variety of Creoles, many of which are more traditional corpora, (e.g., transcriptions of conversations made by linguists with formal training, or scans of documents originally written in the Creole language); though these may not have the relevant annotations for common NLP tasks. Lent et al. (2021) also provides a thorough overview of existing NLP datasets for Haitian Kreyol, Singaporean Colloquial English (Singlish), and Nigerian Pidgin English. In this work, we set about the task of manually verifying each dataset presented by Hagemeijer et al. (2014) and Lent et al. (2021), as well as searching for additional resources. We present all "verified" datasets in Table 1. Here, we use "verified" to mean that we could easily find the resource described in the paper, through either a provided URL in a publication, or through a search engine.

Readers should note that we excluded both extinct Creoles and ostensibly historical Creole data from Table 1. Those interested can see that there are available data for the extinct Virgin Islands Dutch Creole.[5] Other historical Creole data include the Corpus of Mauritian Creole Texts (Baker and Sing, 2007), a collection of texts spanning the 1730 to 1930, and the Surinam Creole Archive (`suca.ruhosting.nl`), which should have historical texts for both Sranan Tongo and Saramaccan, although the hyperlinks are presently broken in this website.

Lastly, to utilize linguistic information about Creoles, the Atlas of Pidgin and Creole Language Structures (APiCS) is an indispensable resource (Michaelis et al., 2013). APiCS is an extension of the popular WALS resource (Dryer and Haspelmath, 2013), but is solely dedicated to pidgins and Creoles.

Unverified Resources Unfortunately many of the Creole corpora reviewed in Hagemeijer et al. (2014) are no longer available, with broken URLs. We describe any resource as "not verifiable", when we cannot track down the resource through the combination of a URL, a simple web search, or through the original publication. These resources may still exist, but they are too difficult to find with a reasonable effort made. The list of "not verifiable" resources can be found in Table 2. We hope that highlighting the "not verifiable" datasets can serve as a call to action in the field, to consider long term data hosting solutions. In order to make the information we gathered about datasets useful in the long-term, we release a community-based webpage.[6] It is hosted on github pages and allows pull requests so that community members can help us maintain up-to-date information about data available for Creoles.

Moreover, in this section, we would also like to discuss book-based corpora. We do not include them in Table 2, as considerable work would need to be done to digitize these datasets, before they can be usable for most NLP tasks. Still, these resources could be useful for those wanting to work with some Creole languages, not listed in Table 1 or Table 2. Creole corpora documented in books include the Corpus of Written British Creole (Sebba, 1998), a corpus of folktales in Tok Pisin (Slone, 2001), and a corpus of Jamaican Creole (Hinrichs, 2006). We also found the following additional resources described by Kouwenberg and Singler (2009) : transcripts of Guyanese Creole (Rickford, 1987), transcripts of English-based Central American Creoles were introduced by (Holm, 1982), and a corpus of various French-based Creoles, such as Louisiana Creole and Reunionese Creole (Corne, 1999).

NLP for Creoles Creole languages, though largely absent from the NLP literature, have been investigated directly in a small number of works. Of the few works actively focused on Creoles, two works explore directly Creole genesis in the context of computational linguistics. First, Daval-Markussen and Bakker (2012) employ phylogenetic tools to explore whether Creole langauges form a unique typological group. By treating each Creole as a list of binary linguistic features, including data from WALS (Dryer and Haspelmath, 2013), they analyze the output of a phylogenetic network program (Huson and Bryant, 2006), to inform their investigation. The overall conclusion made by Daval-Markussen and Bakker (2012), was that Creoles indeed formed their own distinct typological class, distinguishable from non-Creoles. However, this work was later refuted by Murawaki (2016), who argued that the study by Daval-Markussen and Bakker (2012) had some methodological shortcomings. Notably, Murawaki (2016) use APiCS features (Michaelis et al., 2013) to encode Creoles, and utilize different approaches for language evolution modeling, to reach the final conclusion that Creoles are *not* typologically distinct from non-Creole languages.

Meanwhile, Lent et al. (2021) explored the question of how to effectively build language models for three Creole languages (Haitian Kreyol, Singaporean Colloquial English, and Nigerian Pidgin). Their approach involved experimenting with distributionally robust objectives (Oren et al., 2019), to ascertain whether data from a Creole's "parent" languages could help the language model to be more robust. In the end, they found that straightforward training of language models for Creoles, without adding information from their related languages, produced the strongest results, thus highlighting the relative stability of Creoles.

Finally, there have been a handful of other works aiming to develop NLP algorithms usable for end users, primarily in the area of machine translation, for Creoles like Haitian Kreyol, Mauritian Creole, and Nigeran Pidgin (Callison-Burch et al., 2011; Dabre et al.,

---

Language	Resource	Description	Status
Haitian Kreyol	Haitian Disaster Response Corpus (Munro, 2010; Callison-Burch et al., 2011)	SMS	Verified; E-mail authors for access.
	CMU Haitian Corpus `http://www.speech.cs.cmu.edu/haitian/`	Speech and Text Corpora	Verified; E-mail authors for access.
Hawaiian Pidgin	Multilingual Hawai'i Linguistic Landscape Corpus (Purschke, 2021)	Image Repository with Annotations	Verified
Nigerian Pidgin	NaijaSynCor (Bigi et al., 2017)	Speech Recognition	Verified
	JW300 Corpus (Agić and Vulić, 2019)	Parallel Texts for Machine Translation	Verified
	Pidgin UNMT (Ogueji and Ahia, 2019)	Monolingual Texts for Machine Translation	Verified
	Naija-English Codeswitching Corpus (Ndubuisi-Obi et al., 2019)	News Articles with Comments; Annotated for code switching	Verified
	Surface-Syntactic UD Treebank for Naija (Caron et al., 2019)	Universal Dependencies	Verified
	Speech-to-Text Nigerian Pidgin Dataset (Ajisafe et al., 2020)	Speech Recognition	Verified
	NaijaNER (Oyewusi et al., 2021)	Named Entity Recognition	Verified
	Masakhaner (Adelani et al., 2021)	Named Entity Recognition	Verified
	NaijaSenti (Muhammad et al., 2022)	Sentiment Analysis	Verified
Reunionese Creole & Seychellois Creole	Creolica `http://creolica.net/`	Text and Short Stories in HTML or PDFs	Verified
Singlish	National University of Singapore SMS Corpus (Chen and Min-Yen, 2015)	SMS	Verified
	Universal Dependencies for Colloquial Singaporean English (Wang et al., 2017)	UD Treebank	Verified
	Webcrawler for Singaporean Hardware Forum (Tan et al., 2020)	Webcrawler	Verified
Sri Lankan Malay (Endangered)	The Language Archive (Nordhoff, 2006)	Audio and XML	Verified

Table 1: Descriptions of every Creole resource or dataset that we could identify and also verify as being readily available online.

2014; Millour and Fort, 2020; Ahia and Ogueji, 2020).

2.2. Notable Features of Creoles

Many Creole languages are noteworthy for their large capacity for linguistic variation. A speaker's individual style of Creole can vary dramatically depending on social factors, such as their age, ethnicity, geography, and social status. These variations can manifest themselves in different linguistic functions of the Creole, for instance, in the chosen syntax, morphology, or lexical choices (Bajpai et al., 2017). Below, we discuss other features of some (not all) Creoles, that are particularly notable in the context for NLP.

Societal Stigma vs Recognized Status Creole languages are infamously stigmatised (Alleyne, 1971;

Siegel, 1999). To this day, prejudice against Creole languages has thwarted Creole-based education being made available to Creole speakers, for example. The relative status of a language can change drastically, from Creole to Creole. For instance, use of Singlish has been actively discouraged by government officials, citing the need to "Speak Good English".[7] Meanwhile, a handful of other countries have come to embrace Creole (to varying degrees) in their education system, such as Haitian Kreyol, Papiamento, Seychellois Creole, and Tok Pisin (Kouwenberg and Singler, 2009). The relative celebration or suppression a Creole receives will

---

[7]`https://en.wikipedia.org/wiki/Speak_Good_English_Movement`

Language	Resource	Description	Status
Antillean Creole	CREOLORAL `http://ircom.corpus-ir.fr/site/description_projet.php?projet=CREOLORAL`	Audio, Transcriptions, and Translations	Not verifiable
Bastimentos Creole English	Endangered Language Archive `http://elar.soas.ac.uk/deposit/0171`	Audio, Video, Transcriptions, Translations	Not verifiable; Membership required
Gulf of Guinea Creoles	The Gulf of Guinea Creole Corpora (Hagemeijer et al., 2014)	Document Scans and Transcriptions	Limited Verifiability
Haitian Kreyol	Corpus of Northern Haitian Creole `https://www.indiana.edu/~Creole/`	Audio and Transcription	Not verifiable
Malaccan Portuguese Creole	Endangered Language Archive `http://elar.soas.ac.uk/deposit/0123`	Audio, Video, Transcriptions, Translations	Not verifiable; Membership required
Mauritian Creole	ALLEX Project `http://www.edd.uio.no/allex/corpus/africanlang.html`	Concordance of 200k Words	Not verifiable
Nigerian Pidgin	Nigerian Pidgin Tweets (Oyewusi et al., 2020)	Sentiment Analysis	Not Verifiable
Portuguese Creole	CreolData (Schang et al., 2005)	Lexical Database	Not verifiable
Singlish	Singlish Sentiment Lexicon (Bajpai et al., 2017)	Knowledge Base	Not Verifiable
	Singlish SenticNet (Ho et al., 2018)	Sentiment Resource	Not Verifiable

Table 2: Description of Creole datasets presented in our resource survey, which we were not able to verify the existence of. Note here that "Gulf of Guinea Creoles" refers to a collection of four distinct Creole languages: Santome, Angolar, Principense, and Fa d'Ambo.

certainly impact who is speaking the Creole language, and how they will use it.

Spoken Languages Today a large number of Creole languages exist primarily, or almost entirely, as a spoken language only (this can also be a consequence of high stigmatization, as explained in the paragraph above (Sebba, 1997)). If Creole speakers are not typically writing in the language, development of text-based NLP methods may be largely superfluous, unless members of that community have expressed a desire to begin writing (more) in Creole. Consequently, speech technologies may be more relevant to a large number of Creole speaking communities.

Non-standardized Orthography or Grammar Writing conventions for Creoles can vary greatly, from Creole to Creole, and even from speaker to speaker. Given that Creoles arise from a complex process involving several parent languages (Sessarego, 2020), and formal writing education in that Creole is not a guarantee for speakers (Siegel, 1999), there is often no standard way of writing them. On one hand, spellings can depend on an individual and informed by their own oral version of the language (Millour and Fort, 2020). Moreover, spelling and grammar conventions in Creole can also be affected greatly by

a speaker's proficiency in that Creole. For instance, native speakers of Nigerian Pidgin may speak a fluent, fast, and strong variety of the Creole (i.e., less diluted with English), while others speak a weaker Creole, learned as a second language, characterized by heavy use of just one ancestral Nigerian language. This kind of variety in many cases, as with Nigerian Pidgin, is considered a very positive aspect of a Creole, as it grants speakers a lot of opportunity for nuanced expression. Given that contemporary NLP methods are typically not robust to such linguistic variation, it is important not to limit Creole speakers to one register of communication (Doğruöz et al., 2021).

Meanwhile, some Creole languages are undergoing an ongoing cultural shift, with a push towards standardization, in a manner intended to help cultivate a culture of writing in that Creole. For example, in 2014, a language academy was founded for Haitian Kreyol[8],[9]. For those planning to work on text-based Creole applications, it is vital to become attuned to the current writing culture of that Creole's community, and be aware of how speakers are wanting to use their Creole in writing.

---

[8] `https://en.wikipedia.org/wiki/Akademi_Krey%C3%B2l_Ayisyen`
[9] `http://akademikreyol.net/`

Bugs or Features? In summary, many of the features discussed above may be perceived as introducing "challenges" or difficult "problems" for NLP to grapple with, as these features are not shared with high-resource languages, like English or Mandarin. However, these so-called "problems for NLP" are often considered positive features by Creole language speakers themselves. We challenge readers not to think of how they can make Creoles work for NLP, but how NLP can work for Creoles.

3. What's Wanted and What's Needed

In this section, we will give an overview of the key takeaways from our conversations with experts, as well as the major findings from surveying speakers belonging to various Creole speaking communities.

Connecting with Experts As discussed by Bird (2020), building respectful relationships with the relevant community is absolutely necessary, and reaching out to relevant experts is a great first step towards this direction. For the scope of this work, our definition of an "expert" is not strict. We consider an expert to be anyone who is engaged in research, education, or other community outreach, somehow involving the Creole. This can include, for example, individuals working at language schools, field linguists doing research in the area, local scientists in any field, or even graduate students who are native speakers of such languages. Indeed, there are many reasons to begin by reaching out to experts, before even defining your project. First, despite coming from diverse academic backgrounds, experts across different specialities typically speak the same language of scholarship. Although terminological baggage may still interfere with discussion, generally it is easier for fellow field experts to understand, and empathise with each others' goals, than perhaps others. Moreover, even if the experts are not directly working in your field, they may still be familiar or exposed to it. In establishing this relationship, and learning about each other's research or work, there is also a likelihood that some interests overlap, and the opportunity presents itself that you can also help them, which in turn helps to end the norm of treating low-resource language speakers as resources to extract from, and establish a collaborative relationship (Bird, 2020). Additionally, experts also have the authority and knowledge to give you an informed "bird's eye" view of the Creole community, their needs, and desires, as the expert is also a part of it. It's a great (probably necessary) starting point for anyone planning on working on a Creole-language, while not already embedded in the community.

Surveying Creole Speakers With this in mind, discussion with experts alone runs a large risk of missing out on the thoughts of every day Creole speakers, for whom the language technology is ultimately intended. Thus, their thoughts, opinion, desires, and worries are of utmost importance. For this work, we invited Creole speakers to voice their opinions, and to participate in a survey, through both Twitter and Reddit. Two points should be noted about this approach: (1) One limitation of this method is that our posts already unfortunately exclude Creole speakers not also speaking English or French, and (2) We attempt to break away from extractive/exploitative research practices by asking only those with additional interest in the topic to fill out the survey (i.e., for those individuals without substantial interest, we try to minimize the time required for them to contribute to the discussion, by asking general, open ended questions). While the best case scenario would have been to compensate people for their time, as Hu et al. (2011) recall, it can be very difficult to find people willing to participate even for payment. Fortunately for this work, we were still able to find a sizeable number of Creole speakers interested in this topic, and willing to have a discussion with us, even if they did not fill out the survey. For the survey, we had 37 participants in total (35 in English, 2 in French), residing in a diverse range of regions (e.g., Caribbean, Africa, North America, Europe, Asia, and the Pacific). We first asked questions about their linguistic background, and use of various languages in daily life. Then, to target NLP wants and needs, we asked more questions about their language use with regards to technology ("e.g. reading/writing SMS on mobile phone, reading/writing on the internet, reading and writing e-mails, interacting with home assistant devices, etc."). For many questions, we included additional prompts, welcoming participants to expand and explain their answers in short-form, which ultimately yielded many important discussion points from the Creole speakers[10].

For the rest of this section, we will review the consistent themes that arose in our conversations with experts, about the wants and needs of Creole language users. These themes will be further expanded upon by the input provided by Creole speakers from our survey. Again, not all themes will be relevant for every Creole. On the contrary, themes seem to be primarily relevant to Creoles with very specific attributes in common (see Section 4).

Is Language Technology Wanted or Needed? As discussed throughout this paper, Creole languages are incredibly diverse, including in the way people want (or don't want) to use these languages to interact with technology. Thus, it should come as little surprise, that the answer to the question: "Is language technology wanted and/or needed for this language?", can be everything from "Yes!", "Some technology would be nice", "No", and "Why would you waste time doing that?!", among others.

Amongst both experts and Creole speaking survey respondents, the answer to this question appeared to be largely contingent on how proficient members of the larger community are in the local, high-prestige lan-

---

[10]Please contact us directly if you would like access to our surveys.
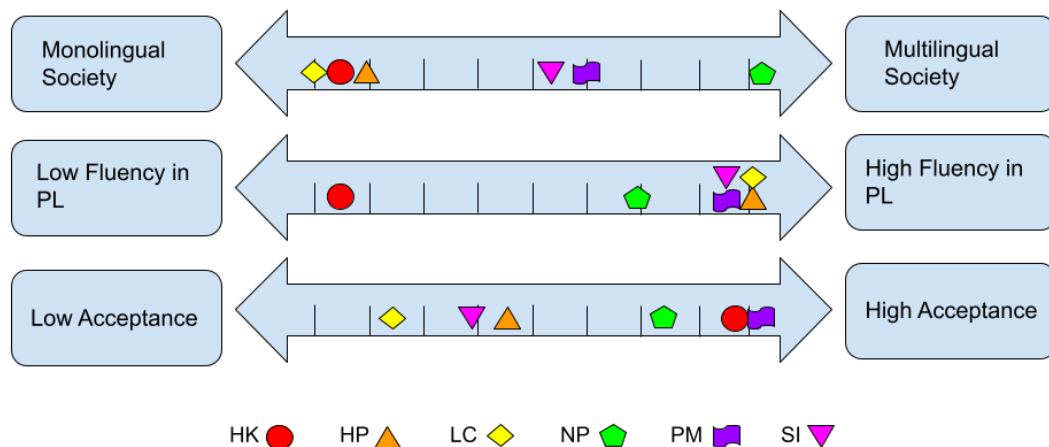
Figure 1: We map a sample of Creole languages to our proposed Creole continuum for language technology. PL here refers to "Prestige Language". We map Haitian Kreyol (HK), Hawaiian Pidgin (HP), Louisiana Creole (LC), Nigerian Pidgin (NP), Papiamento (PM), Singlish (SI).

guage (typically English, French, or Portuguese). For instance, a limited subset of the population of Haiti speaks French, and thus Haitian Kreyol is used in most aspects of every day life, and technology to ease the use of Haitian Kreyol is highly desired. On the other end of the spectrum, experts and speakers of Hawaiian Pidgin had difficulties coming up with reasons why language technology support for their Creole would be particularly useful, as the overwhelming majority of speakers (if not all) are highly proficient in English.

Current Obstacles In our discussions, some expressed that they already use their Creole for basic tasks, such as texting friends, but that it was not always easy. For example, existing autocomplete or autocorrect software on phones and computers (installed in the relevant high-prestige language, as these technologies are not readily available to Creoles) often automatically "corrects" Creole spellings or words, and inadvertently suppressing written Creole usage in daily life.

Another issue that Creole speakers mentioned about existing speech technology, was the lack of support for Creole accents or casual code-switching with commonplace Creole words. For instance, navigational assistants for GPS struggle to understand Hawaiian Pidgin accents, in addition to being unable to pronounce local street names, which can be uttered in Hawaiian Pidgin, but not in Standard American English. Extending existing speech technology for dominant, high-prestige languages in this space is much desired, and can be preferable over having a separate, Creole-only system. But without these modifications, existing language technology for the local, high-prestige language can actively harm Creole speakers.

Speech Technology As discussed in the Background, many Creole languages are used almost exclusively used in spoken conversation. For such Creoles, text-based language technology are likely moot. Although this can change with time, we encourage readers interested in working in Creole spaces to check with experts and communities, to ascertain if text-based technologies are even needed.

When speech technology was discussed, most Creole languages expressed interest and desire in having speech technology (both text-to-speech and speech-to-text), with the small exception of Creole languages under threat of decreolization (the process by which a Creole ceases to exist), where language revitalization is the dominant concern. But overall, speech technology was perceived by experts and survey respondents to be the most desirable and wanted language technology.

Facilitating Writing Some Creole speaking communities already do a lot of writing in their Creole language (despite some obstacles, as we have seen), and/or are trying to foster a culture of writing in the Creole, including standardizing the language. In our conversations with experts and survey responders, we note that there is an expressed need by some Creole communities for basic word processing tools, such as word processors, spell-checkers, grammar-checkers, auto-transcription, etc. However, we found that not all Creole communities welcome *all* of these technologies equally. For example, speakers of Haitian Kreyol mostly welcome spell-checkers, meanwhile speakers of Nigerian Pidgin would eschew these, as it constrains their language use. This point demonstrates how, even when there is a shared desire for a specific kind of language technology, the implementation and specific

needs for a Creole can be highly specialized. Lastly, we note that, just as you must learn to walk before you can run, technologies that ease or improve writing in Creoles may be necessary before Creole speakers could have a need for semantic parsing, for example.

Question Answering and Machine Translation Both question answering (QA) and machine translation (MT) came up as desired technologies for many Creoles, albeit for different reasons. For Creoles already used online to some extent, QA could improve online search, while MT from Creole into a high resource language, or vice-versa, could provide access to other parts of the world for Creole speakers. Also, MT was cited as desirable for even some endangered Creoles, like Louisiana Creole, as it could help with revitalization. For example, automatic translation from English or French to Louisiana Creole, could allow people to enjoy new domains in Louisiana Creole, and in turn assist with (re)learning the language.

Summary Overall, our discussions with Creole experts and every day Creole speakers underscored how diverse the needs of Creoles can be, for even within one group of languages. We hope this discussion, and the themes put forward, can serve as a springboard for those planning to work on NLP for Creoles.

4. Creole Continuum for Language Technology

While the previous section demonstrated that Creoles are not a monolith when it comes to wants and needs for language technology, we did observe several patterns, where Creoles seemed to cluster together, based on their language technology needs, depending on a few shared attributes. To this effect, we introduce a Creole continuum for language Technology (inspired by the post-Creole continuum (DeCamp, 1971)), and propose that there are three key factors that can heavily influence the general needs of a Creole, as follows: (1) Monolingual, Bilingual, or Multilingual community (in other words, is the Creole a lingua franca, facilitating cross-lingual communication?); and (2) General fluency in the relevant prestige language (i.e., do most people also speak the more globally prestigious language, and get on fine, without the Creole?); and (3) Societal acceptance of Creole (e.g., is the Creole language embraced by society as large, or does the Creole struggle from a bad reputation?). We present this continuum in Figure 1, with a small collection of Creole languages, to serve as an example.[11]

The first pattern we would like to draw to the reader's attention to is that the Creoles existing within predominantly monolingual societies, that are also highly fluent in the local prestige language, are those that do not

---

[11]We specifically intend the graph axis to be flexible for interpretation, as different Creoles will have different needs, and strict or concrete axis categories may risk reinforcing existing marginalization.

need language technology (Hawaiian Pidgin), at least not beyond revitalization (Louisiana Creole). Also, any time that the larger society exhibits very low fluency in the prestige language, language technology is much more likely to be wanted and needed in these communities (Haitian Kreyol). For other Creoles, it is not so clear cut, though. For example, both Singlish and Papiamento exist in generally multilingual societies, with a majority of speakers also fluent in Dutch and English, respectively, and yet the increased societal acceptance of Papiamento (Papiamento is a recognized language of Aruba), means that speakers are more likely to welcome or express needs for language technologies. Still, Singlish is not to be completely neglected, but due to its lower acceptance, language technology suiting more informal situations (e.g. dialog) will likely be more relevant. And finally, the speakers of languages with high acceptance of Creole, namely Papiamento, Haitian Kreyol, and Nigerian Pidgin, are those who typically have the most clear cut wants and needs from language technology, as they already likely use their Creole to interact with technology.

5. Conclusion

In this work, we have demonstrated that Creole languages should be of larger interest to the NLP community, and we provide a survey of resources and NLP research produced for Creoles. In doing this, we have also shown that Creoles cannot be conflated together, if we are to make language technology that is truly useful for a community. Truly, the best approach to developing NLP for Creoles to to get in contact with both experts and community members, and listen earnestly to their wants and needs for language technologies, as well as what is specifically not wanted.

6. Acknowledgements

7. Bibliographical References

Adelani, D. I., Abbott, J., Neubig, G., D'souza, D., Kreutzer, J., Lignos, C., Palen-Michel, C., Buzaaba, H., Rijhwani, S., Ruder, S., Mayhew, S., Azime, I. A., Muhammad, S. H., Emezue, C. C.,

Nakatumba-Nabende, J., Ogayo, P., Anuoluwapo, A., Gitau, C., Mbaye, D., Alabi, J., Yimam, S. M., Gwadabe, T. R., Ezeani, I., Niyongabo, R. A., Mukiibi, J., Otiende, V., Orife, I., David, D., Ngom, S., Adewumi, T., Rayson, P., Adeyemi, M., Muriuki, G., Anebi, E., Chukwuneke, C., Odu, N., Wairagala, E. P., Oyerinde, S., Siro, C., Bateesa, T. S., Oloyede, T., Wambui, Y., Akinode, V., Nabagereka, D., Katusiime, M., Awokoya, A., MBOUP, M., Gebreyohannes, D., Tilaye, H., Nwaike, K., Wolde, D., Faye, A., Sibanda, B., Ahia, O., Dossou, B. F. P., Ogueji, K., DIOP, T. I., Diallo, A., Akinfaderin, A., Marengereke, T., and Osei, S. (2021). MasakhaNER: Named entity recognition for African languages. *Transactions of the Association for Computational Linguistics*, 9:1116–1131.

Agić, Ž. and Vulić, I. (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, July. Association for Computational Linguistics.

Ahia, O. and Ogueji, K. (2020). Towards supervised and unsupervised neural machine translation baselines for nigerian pidgin. *ArXiv*, abs/2003.12660.

Ajisafe, D., Adegboro, O. G., Oduntan, E., and Arulogun, T. O. (2020). Towards end-to-end training of automatic speech recognition for nigerian pidgin. *ArXiv*, abs/2010.11123.

Alleyne, M. (1971). Acculturation and the cultural matrix of creolization. *Pidginization and*, pages 169–186.

Artetxe, M., Labaka, G., and Agirre, E. (2020a). Translation artifacts in cross-lingual transfer learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7674–7684, Online, November. Association for Computational Linguistics.

Artetxe, M., Ruder, S., and Yogatama, D. (2020b). On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online, July. Association for Computational Linguistics.

Bajpai, R., Poria, S., Ho, D., and Cambria, E. (2017). Developing a concept-level knowledge base for sentiment analysis in singlish. *CoRR*, abs/1707.04408.

Baker, P. and Sing, G. F. (2007). *The making of Mauritian Creole. Analyses diachroniques à partir des textes anciens*. Number 9. Battlebridge.

Bickerton, D. (1984). The language bioprogram hypothesis. *Behavioral and brain sciences*, 7(2):173–188.

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

Bird, S. (2020). Decolonising speech and language technology. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3504–3519, Barcelona, Spain (Online), December. International Committee on Computational Linguistics.

Budur, E., Özçelik, R., Gungor, T., and Potts, C. (2020). Data and Representation for Turkish Natural Language Inference. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8253–8267, Online, November. Association for Computational Linguistics.

Chris Callison-Burch, et al., editors. (2011). *Proceedings of the Sixth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Edinburgh, Scotland, July.

Caron, B., Courtin, M., Gerdes, K., and Kahane, S. (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, August. Association for Computational Linguistics.

Chen, T. and Min-Yen, K. (2015). The national university of singapore sms corpus.

Conneau, A., Rinott, R., Lample, G., Williams, A., Bowman, S., Schwenk, H., and Stoyanov, V. (2018). XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium, October-November. Association for Computational Linguistics.

Corne, C. (1999). *From French to Creole: The development of new vernaculars in the French colonial world*, volume 5. Westminster creolistics.

Dabre, R., Sukhoo, A., and Bhattacharyya, P. (2014). Anou tradir: Experiences in building statistical machine translation systems for mauritian languages – creole, English, French. In *Proceedings of the 11th International Conference on Natural Language Processing*, pages 82–88, Goa, India, December. NLP Association of India.

Daval-Markussen, A. and Bakker, P. (2012). Explorations in creole research with phylogenetic tools. In *EACL 2012*.

DeCamp, D. (1971). *Toward a generative analysis of a post-creole speech continuum*. Cambridge University Press.

DeGraff, M. (2001). On the origin of creoles: A cartesian critique of neo-darwinian linguistics. *Linguistic Typology*, 5(2/3):213–310.

DeGraff, M. (2003). Against creole exceptionalism. *Language*, 79(2):391–410.

DeGraff, M. (2005). Linguists' most dangerous myth: The fallacy of creole exceptionalism. *Language in society*, 34(4):533–591.

Doğruöz, A. S., Sitaram, S., Bullock, B. E., and Toribio, A. J. (2021). A survey of code-switching: Linguistic and social perspectives for language tech-

nologies. 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, August. Association for Computational Linguistics.

Matthew S. Dryer et al., editors. (2013). *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Graham, Y., Haddow, B., and Koehn, P. (2020). Statistical power and translationese in machine translation evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 72–81, Online, November. Association for Computational Linguistics.

Hagemeijer, T., Généreux, M., Hendrickx, I., Mendes, A., Tiny, A., and Zamora, A. (2014). The Gulf of Guinea creole corpora. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 523–529, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

Henri, F., Stump, G., and Tribout, D. (2020). Derivation and the morphological complexity of three french-based creoles. In *The Complexities of Morphology*, pages 105–135. Oxford University Press.

Hinrichs, L. (2006). *Codeswitching on the Web*. John Benjamins Amsterdam.

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

Holm, J. (1982). *Central American English*, volume 2. John Benjamins Publishing.

Hu, C., Resnik, P., Kronrod, Y., Eidelman, V., Buzek, O., and Bederson, B. B. (2011). 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, July. Association for Computational Linguistics.

Huson, D. H. and Bryant, D. (2006). Application of phylogenetic networks in evolutionary studies. *Molecular biology and evolution*, 23(2):254–267.

Joshi, P. M., Santy, S., Budhiraja, A., Bali, K., and Choudhury, M. (2020). The state and fate of linguistic diversity and inclusion in the nlp world. In *ACL*.

Kouwenberg, S. and Singler, J. V. (2009). *The handbook of pidgin and creole studies*. John Wiley & Sons.

Lent, H., Bugliarello, E., de Lhoneux, M., Qiu, C., and Søgaard, A. (2021). On language models for creoles. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages

58–71, Online, November. Association for Computational Linguistics.

Liu, F., Bugliarello, E., Ponti, E., Reddy, S., Collier, N., and Elliott, D. (2021). Visually grounded reasoning across languages and cultures. *ArXiv*, abs/2109.13238.

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

Susanne Maria Michaelis, et al., editors. (2013). *APiCS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Millour, A. and Fort, K. (2020). Text corpora and the challenge of newly written languages. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 111–120, Marseille, France, May. European Language Resources association.

Mirzakhalov, J., Babu, A., Ataman, D., Kariev, S., Tyers, F., Abduraufov, O., Hajili, M., Ivanova, S., Khaytbaev, A., Laverghetta Jr., A., Moydinboyev, B., Onal, E., Pulatova, S., Wahab, A., Firat, O., and Chellappan, S. (2021). A large-scale study of machine translation in Turkic languages. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5876–5890, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Muhammad, S. H., Adelani, D. I., Ruder, S., Ahmad, I. S., Abdulmumin, I., Bello, B. S., Choudhury, M., Emezue, C. C., Abdullahi, S. S., Aremu, A., Jeorge, A., and Brazdil, P. (2022). Naijasenti: A nigerian twitter sentiment corpus for multilingual sentiment analysis.

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

Murawaki, Y. (2016). Statistical modeling of creole genesis. In *NAACL*.

Muysken, P. and Smith, N. (1986). *Substrata versus universals in creole genesis: papers from the Amsterdam Creole Workshop, April 1985*, volume 1. John Benjamins Publishing.

Ndubuisi-Obi, I., Ghosh, S., and Jurgens, D. (2019). Wétin dey with these comments? modeling sociolinguistic factors affecting code-switching behavior, in nigerian online discussions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*.

Nekoto, W., Marivate, V., Matsila, T., Fasubaa, T., Fagbohungbe, T., Akinola, S. O., Muhammad, S., Kabongo Kabenamualu, S., Osei, S., Sackey, F., Niyongabo, R. A., Macharm, R., Ogayo, P., Ahia, O., Berhe, M. M., Adeyemi, M., Mokgesi-Selinga, M., Okegbemi, L., Martinus, L., Tajudeen, K.,

Degila, K., Ogueji, K., Siminyu, K., Kreutzer, J., Webster, J., Ali, J. T., Abbott, J., Orife, I., Ezeani, I., Dangana, I. A., Kamper, H., Elsahar, H., Duru, G., Kioko, G., Espoir, M., van Biljon, E., Whitenack, D., Onyefuluchi, C., Emezue, C. C., Dossou, B. F. P., Sibanda, B., Bassey, B., Olabiyi, A., Ramkilowan, A., Öktem, A., Akinfaderin, A., and Bashir, A. (2020). Participatory research for low-resourced machine translation: A case study in African languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2144–2160, Online, November. Association for Computational Linguistics.

Ogueji, K. and Ahia, O. (2019). Pidginunmt: Unsupervised neural machine translation from west african pidgin to english. *ArXiv*, abs/1912.03444.

Ogueji, K., Zhu, Y., and Lin, J. (2021). Small data? no problem! exploring the viability of pretrained multilingual language models for low-resourced languages. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 116–126, Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Oren, Y., Sagawa, S., Hashimoto, T. B., and Liang, P. (2019). Distributionally robust language modeling. *arXiv preprint arXiv:1909.02060*.

Oyewusi, W. F., Adekanmbi, O., and Akinsande, O. (2020). Semantic enrichment of nigerian pidgin english for contextual sentiment classification. *ArXiv*, abs/2003.12450.

Oyewusi, W. F., Adekanmbi, O., Okoh, I., Onuigwe, V., Salami, M. I., Osakuade, O., Ibejih, S., and Musa, U. A. (2021). Naijaner : Comprehensive named entity recognition for 5 nigerian languages. *ArXiv*, abs/2105.00810.

Parkvall, M. et al. (2008). The simplicity of creoles in a cross-linguistic perspective. *Language complexity: Typology, contact, change*, pages 265–285.

Purschke, C. (2021). Crowdscapes. participatory research and the collaborative (re) construction of linguistic landscapes with lingscape. *Linguistics Vanguard*, 7(s1).

Rickford, J. R. (1987). *Dimensions of a Creole continuum: History, texts & linguistic analysis of Guyanese Creole*. Stanford University Press.

Schang, E., Rougé, J.-L., Eshkol, I., and Petit, M. (2005). Creoldata: A lexical database on creole languages. *Revue francaise de linguistique appliquee*, 10(1):65–76.

Sebba, M. (1997). *Contact languages: Pidgins and creoles*. Macmillan International Higher Education.

Sebba, M. (1998). Phonology meets ideology: the meaning of orthographic practices in british creole. *Language problems and language planning*, 22(1):19–47.

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

Siegel, J. (1999). Stigmatized and standardized varieties in the classroom: Interference or separation? *Tesol Quarterly*, 33(4):701–728.

Slone, T. H. (2001). *One Thousand One Papua New Guinean Nights: Tales form 1972-1985*, volume 1. Masalai Press.

Tan, S., Joty, S., Varshney, L., and Kan, M.-Y. (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, November. Association for Computational Linguistics.

Thomason, S. G. and Kaufman, T. (1992). *Language contact, creolization, and genetic linguistics*. Univ of California Press.

Volansky, V., Ordan, N., and Wintner, S. (2013). On the features of translationese. *Digital Scholarship in the Humanities*, 30(1):98–118, 07.

Wang, H., Zhang, Y., Chan, G. L., Yang, J., and Chieu, H. L. (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, July. Association for Computational Linguistics.

8. Language Resource References

Nordhoff, Sebastian. (2006). *Collection 'Sri Lanka Malay'*. The Language archive.