Molyé: A Corpus-based Approach to Language Contact in Colonial France

Rasul Dent, Juliette Janes, Thibault Clérice, Pedro Ortiz Suarez, Benoît Sagot

Molyé: A Corpus-based Approach to Language Contact in Colonial France

Rasul Dent[1], **Juliette Janès**[1], **Thibault Clérice**[1], **Pedro Ortiz Suarez**[2], **Benoît Sagot**[1]
[1]Inria, Paris, {firstname.lastname}@inria.fr
[2]Common Crawl Foundation, Paris, pedro@commoncrawl.org

Abstract

Whether or not several Creole languages which developed during the early modern period can be considered genetic descendants of European languages has been the subject of intense debate. This is in large part due to the absence of evidence of intermediate forms. This work introduces a new open corpus, the Molyé corpus, which combines stereotypical representations of three kinds of language variation in Europe with early attestations of French-based Creole languages across a period of 400 years. It is intended to facilitate future research on the continuity between contact situations in Europe and Creolophone (former) colonies.

1 Introduction

Between the 15th and 19th centuries, several languages developed in colonized territories, which, while sharing a large amount of vocabulary with existing European languages, differ considerably in morphology and syntax. These languages are often labeled English-based [or lexified] Creoles, French-based Creoles, Portuguese-based Creoles, etc., according to the language they share most of their vocabulary with, which is itself called the lexifier. One long standing question has been why the grammars of these languages diverged from their lexifiers to a greater extent than the vocabulary (de Sousa et al., 2019). Much of the difficulty in answering this question stems from harsh social conditions discouraging linguistic documentation and environmental conditions destroying much of what had been documented (McWhorter, 2018).

For French-based Creole languages (FBCLs), which developed on islands and isolated continental settlements during the 17th and 18th centuries (Chaudenson, 2001) [1], reliable documentation largely dates from the mid-late 18th century onward (Hazaël-Massieux, 2008). However, we note that the formative period of FBCLs coincided with a period of French political and cultural dominance and extensive literary production known as the Grand Siècle. The cultural works of the period are replete with numerous stereotypes of the speech of several social groups, such as urbanized peasants and Swiss soldiers. Despite various issues detailed by Ayres-Bennett (2000), these representations are relevant for FBCLs insofar as they demonstrate several interesting morphosyntactic developments.

Here, we introduce the Molyé corpus, which regroups 68 works that contain examples of either the aforementioned stereotypes or early attestations of FBCLs. This list has been curated from a larger collection of 301 documents identified at the time of publication. [2] We begin by giving an overview of related corpora and how we approach historical linguistics as an instance of multi-label language identification. After giving some linguistic context, we also explain the process of identifying Creole-like features in French literary works, encoding said works into XML-TEI, and then compiling groups of quotes into a timeline. Finally, we present summary statistics and conclude by giving examples of how our corpus highlights intra-European contact.

2 Related Work

In recent years, Creole languages have garnered attention in the field of natural language processing. On the one hand, Lent et al. (2022b) have explored how these languages challenge the assumed desirability of certain applications. On the other hand, Lent et al. (2022a) and Robinson et al. (2023) argue that language models for concrete problems may shed light on theoretical issues as well. Simultaneously, in computational historical linguistics, List (2024) has declared the inferral of morpheme boundaries and the detection of layers of language contact to be major open problems. Our work addresses both the paucity of early Cre-

[1]Except Tayo in 19th century New Caledonia.

[2]The corpus can be accessed and downloaded at the following address: https://github.com/defi-colaf/Molye.

ole documentation and the issue of multiple layers of language contact through the applied lens of language identification.

2.1 (Digital) Diachronic Corpora

For several Creolophone regions, such as Louisiana (Neumann-Holzschuh, 1987), the Caribbean (Hazaël-Massieux, 2008), Réunion (Chaudenson, 1981), and Mauritius (Baker et al., 2007; Chaudenson, 1981), diachronic corpora have been compiled in print. However, to our knowledge, only the Mauritian corpus has been systematically digitized and made readily accessible (Fon Sing, 2013). Beyond this, certain historical works have been digitized for inclusion in analysis-oriented private diachronic corpora (Mayeux, 2019), or for applied goals like machine translation (Robinson et al., 2024), and others have been individually published by groups such as the European Group for Research on Creole Languages (Hazaël-Massieux, 2013a,b).

To digitize documents in a way that can facilitate reuse, we rely on the standards of Text Encoding Initiative (TEI) (TEI Consortium eds., 2023). Adherence to these guidelines has produced diachronic corpora which span several centuries, such as Favaro et al. (2022). For the languages of France, Bermudez Sabel et al. (2022) have addressed some of the challenges of building comparable corpora for the parent-daughter pair of Latin and French. Similarly, Ruiz Fabo et al. (2020) explore how digitizing 19th century Alsatian theatre aids sociolinguistic studies.

2.2 Multi-Label Language Identification

Algorithms for determining the language of a given text generally rely on tokenizing the text and comparing the tokens against a learned (or explicitly defined) representation of a language (Jauhiainen et al., 2019). For analytic languages written in the Latin alphabet (i.e. FBCLs), tokens generally align with either words or letters. With closely-related languages, there is sometimes only a difference of a singular word or even letter between one variety and another, even in longer documents (Ljubesic et al., 2007; Caswell et al., 2020). In these cases, we can specify disjunctive features such as words/phrases that are thought to separate the varieties to either affirm or reject a label. In the absence of such features, the same string may be valid in multiple languages, which can make it more accurate to assign multiple language labels to the same string (Bernier-Colborne et al., 2023; Keleg and Magdy, 2023).

3 Linguistic Background

The backbone of our corpus is applying multi-label language identification based on disjunctive features across time. In concrete terms, we sought out distinctly "Creole" features in Europe before and during the colonial expansion. As such, we briefly review a few characteristics of FBCLs, followed by French literary stereotypes.

3.1 French-based Creole Languages

3.1.1 Description

While the notion that all Creoles can be defined in purely linguistic terms, as explored by McWhorter (1998); Bakker et al. (2011), is controversial, FBCLs are agreed to share several traits which distinguish them from standard French. Firstly, they generalized the use of tonic pronouns in places where the latter would use weak clitic pronouns (Syea, 2017). In cases where French does not have a weak pronoun (i.e. 'nous'), they still differ by not allowing preverbal cliticization of object pronouns. Additionally, while French relies on a system of fusional conjugations, where verb endings mark person, number, tense, aspect and in the case of the past participle, gender, at the same time, FBCLs add person-invariant combinations of Tense-Aspect-Mood (TAM) markers (Syea, 2017; Baker and Corne, 1982). These differences are demonstrated by the anteriority marker 'té', and the conditional marker 'sré' in the phrase 'Pour sûr si vou **té** capab changé vou lapo pou so kenne, vou **sré** pa di non' (Mercier, 1881) [3]. Furthermore, FBCLs do not have an explicit copula in several structures where one is required in French (and English), as demonstrated by the phrases 'Comme **vous bel**' [4] and 'vou **papa riche**' [5] in Figure 2.

3.1.2 Theories of Origins

As previously stated, the relationship of Creole languages to lexifiers remains a topic of intense debate. For this work, one relevant hypothesis, as explored by Chaudenson (2001), suggests that the accumulation of the defining characteristics occurred over several waves of second language acquisition, as opposed to being the result of a complete break in

[3] Surely, if you could trade your skin for his/hers, you would not say no.

[4] how **you** [are] **beautiful**.

[5] you[r] **dad** [is] **rich**.

transmission of syntax, as suggested by McWhorter (2018) and Thomason and Kaufman (1988). Another line of inquiry explores the extent to which "foreigner talk", which is to say a particular kind of simplified register that people adopt when they feel their interlocutors do not have sufficient competence in the language, may have contributed to certain developments in Creole morphology and syntax (Ferguson, 1981, 1975). For Portuguese- and Spanish-based Creoles, there is a long history of triangulating Iberian versions of foreigner talk with early modern literary stereotypes and contemporary Afro-Hispanic varieties to get an idea of the range of linguistic variation in the early modern Iberian empires (Kihm, 2018; Lipski, 2001). In the following section, we explore how a similar approach can applied to French.

3.2 French Literary Stereotypes

Up to the 20th century, most people in France spoke regional languages (Lodge, 2003). In the Northern half of mainland France, most of these languages are part of the Oïl dialect continuum, which is itself part of a larger Western Romance continuum. However, non-Romance languages such as Breton (North-West) and Flemish (North) are spoken as well. From the Middle Ages on a particular Oïl variety, associated with prestigious actors was gradually codified into the standard language of the Kingdom of France. This variety was also adopted as a *lingua franca* throughout Europe, as an alternative to Latin. During the 17th and 18th centuries, the process of codification culminated in a well delimited variety known as Classical French.

However, the codified "bon usage", was not the only supralocal speech used in France. Even within the Paris region, there was a great deal of variation within what could be considered "French" (Wittmann, 1995). In broad terms, we distinguish three types of variation: dialectal and sociolectal variation from the Oïl domain, standard French with regional accents, and interlanguages, especially from L1 speakers of Germanic languages [6]. In all three of these cases, we find stereotyped combinations of a finite number of highly stigmatized features in a variety of works, including plays, novels, songs, and personal letters.

[6]Other phenomena, such as the mix of various forms of Occitan in *Monsieur de Pourceaugnac* described by Sauzet and Brun-Trigaud (2015), are beyond our immediate scope.

3.2.1 Peasant French

By the early 1600s, several features of rural usage in the outskirts of Paris (and Western France), such as the combination of clitic pronoun 'je' with the plural affix '-ons', were developed into a convention for representing lower class characters in literature (Lodge, 1991; Ayres-Bennett, 2004), as seen in this example from *La Mère confidente* (Marivaux, 1735): '**Je** sav**ons** bian ce que c'est; **j'ons** la pareille.'[7] Although this stereotype was frozen relatively early on, the highlighted combination was used in France and its colonies throughout the colonial period and still exists in Acadian French in particular, albeit more commonly as a plural form (King et al., 2004).

3.2.2 Gascon Accent

French also came to be spoken as a second language in areas where the regional languages were even more different from French. In these case, the native languages had some influence on pronunciation. In classical French theatre, one common stereotype of such regional pronunciation is the Gascon accent, which can be identified through its betacism (conflating b and v) and fronting of the schwa (replacing e with é). The character Fontignac from *L'île de la raison* (Marivaux, 1727) demonstrates the convention with this line: '...**b**ous m**é** d**é**mandez c**é** qu**é** **b**ous êtes ; mais j**é** n**é** **b**ous **b**ois pas ; mettez-**b**ous dans un microscope.'[8]

3.2.3 Germanic Baragouin

Germanic Baragouin [9] (henceforth just Baragouin) is our name for a group of stereotypes which simultaneously combine traits of foreigner talk, foreign accents, and Oïl dia- and sociolectal variation. In the early modern period, there are two main variations: the Anglo-Baragouin attributed to English (and Scots) speakers, and Continental Baragouin associated with German and Dutch, and more specifically, Swiss and Flemish speakers (Leach, 2020; Damm, 1911). A third, industrial-era Flemish Baragouin also developed around the turn of the 20th century in the cities of Tourcoing and Roubaix near the French-Belgian border (Landrecies, 2001). The main differences between

[7]We/I know what [the task] is, we/I have a similar one.

[8]... You ask me what you are; yet I do not see you. Put yourself in a microscope.

[9]The word "baragouin" [gibberish] was also used to describe a variety of contact phenomena ranging from accented pronunciation to genuine pidgins like that used with the Caribs in the Lesser Antilles (Wylie, 1995).

these sub-groups of Baragouin lie in phonetics. The Continental Baragouin generalizes final-consonant devoicing into a complete neutralization of several consonant pairs, such as /b/-/p/, /k/-/g/, /v/-/f/ and /t/-/d/. Similarly, the industrial-era Flemish version features palatal fronting of /ʃ/ and /ʒ/ to /s/ and /z/. These traits are mostly absent in the English version.

In terms of morphosyntax, Baragouin shares some traits with Creoles, such as the generalization of strong pronouns, weakening of grammatical gender, and reduced verbal inflection (Haas, 2015). However, Baragouin also retains an overt copula and systematically inserts third-person pronouns before verbs, which results in sentences such as '**Toi *li* être**, par m*on* foi, la plus *p*elle meilleure h*i*meur du monde [10]' (Guelette, 1740). The latter features have a special importance, which we explore further in Section 7.1.

4 Corpus Creation

The compilation of the corpus was realized in three overlapping phases. During the first phase, we identified documents which contained n-grams thought to be highly disjunctive between French and various FBCLs. After identifying the documents the next step was to convert them relevant samples into the XML-TEI schema of a broader project. Lastly, we classified the documents by location and period and extracted the relevant quotes into a combined XML document to facilitate the preliminary analysis presented in Section 6.

4.1 Document identification

The basic strategy was to search Gallica, the digitized library of the Bibliothèque nationale de France [11], Delpher, its Dutch equivalent, and later Google Books for disjunctive n-grams. Examples include monograms (e.g. 'mo', 'to', 'yé'), bigrams, e.g ('mo(n) femme','mo(n) z'enfant'), and higher n-grams. Due to variation in both French orthography and the conventions/contact varieties themselves, an iterative approach was taken, with documents collected on the first pass providing more "unusual" n-grams for subsequent searches. In the earliest stages, we did not note the exact searches, but later began to record the search terms as well. In a later stage, we also added several Creole sources known through secondary literature in order to facilitate in-depth diachronic comparison.

Because we are working with stereotypes, a certain level of similarity was to be expected. Nevertheless, in some cases, we found that certain works go into the realm of explicit reference and/or pastiche. For direct quotation, there is *Les fêtes de l'amour et de Bacchus* which includes a reprise of the linguistic humor from *Le Bourgeois Gentilhomme*, among other pieces. As far as pastiche, we can highlight the early 16th century *Testament du Gentil Cossoys* and its early 17th century reprise, the *Testament d'vn Escossois*. The latter is a simultaneously condensed and updated version of the former. Thus where the original reads 'Adiou par tout nobe royaulm de Frans / Adiou comman le povre pals de Cos...'[12] (Smith, 1920), the reprisal has ' Ady par tout le Royaume de France/ Premierement ady le pay de Coss... [13]' (Sigogne, 1620)

Search	Lang Type	Document	Year
"ly va"	Baragouin	Francion	1630
li-même	Peasant	L'Épreuve	1740
conné li	L. Creole	L'autre monde	1855

Table 1: Sample Searches and Documents

4.2 Encoding Documents

Given both the large number of documents it was necessary to establish an order of priority for incorporating works into the corpus. We initially focused on both Baragouin and Peasant French in works of classical theatre that had already encoded by sources such as theatre-classique.fr (Fièvre, 2007). Beyond the core of classical French theatre, however, a wide variety of genres are represented. These include poetry, songs, religious material, short prose, and an entire novel. The subject matter exhibits a similar degree of variability. In the Baragouin section alone, we find, among other things, two mock-testaments, a criticism of military leadership, a love letter, and a discussion about the political implications of an ongoing civil war.

After treating the extant XML, we explored semi-automatic generation of XML-TEI documents from semi-structured sources such as Wikisource, as

[10] By my faith, you are [lit.**you *it* be**] the most beautiful best humour of the world

[11] National Library of France

[12] Adieu to all noble kingdom of France / Adieu likewise poor Scotland

[13] Adieu to all the Kingdom of France/ Firstly adieu Scotland

well as directly from scanned documents. In the former case, we used relatively simple custom Python scripts to facilitate conversion to TEI, such as wrapping all of the lines in a ‹p› (paragraph) or ‹l› (line/verse) tag, and then identifying divisions and headers manually. In the the latter case, this involved a considerable amount of manual transcription due to the diversity of genres and formats. For shorter works, such as poems and songs, we used eScriptorium (Kiessling et al., 2019) to perform text recognition with the CATMuS Print model (Gabay et al., 2024). However, more complex layout (e.g. newspaper) were transcribed manually. For longer works, we entered the relevant quotes directly into a file of excerpts.

4.3 Linguistic Annotation

Since this corpus is in large part intended to illustrate a sociolinguistic continuum assigning discrete linguistic labels poses distinct challenges. Although it is clearly anachronistic to speak of "[Colony] French/Creole" before the founding of a given colony, we observe that in certain cases, namely in Réunion and Louisiana, the "approximative French", "pidginized French", or "pre-Creole" (depending on one's point of view) bears striking continuity with Baragouin at the morphological and syntactic levels. In a parallel fashion, early texts which are clearly "Creole", such as "La passion de Notre Seigneur selon St Jean en Langage Negre", display combinations of features which make it difficult to say *which* Creole based on purely linguistic data.

Following the brief outline given in 3, we distinguish between five main kinds of language: Classical French (met-fr), Peasant French (fra-dia), (Gascon) Accented French (fra-gsc), Baragouin (subdivided into fra-ang, fra-deu, and fra-nld), and (pre)-Creoles. The Creole portion is in turn subdivided into four regions and labelled using the respective ISO codes: Réunion (rcf), Louisiana (lou), Haitian (hat), and French Guianese (gcr). For the initial work, we have somewhat simplified the question of diachronic and dialectal continua by assigning one label based on the territory a document claims (or has been presumed) to represent, with the exception of grouping the earlier "Flemish" baragouin with the German one rather than the later Flemish Baragouin, based on the differences described in Section 3.2.3.

For adding linguistic labels to documents, we

```
<div type="scene" n="10">
    ...
    <sp who="JACQUES" xml:lang="mau">
        <speaker>JACQUES.</speaker>
        <p>... Enfin pourtant , li jetté son zépée ,
            li remetté pistolet dans son place ,
            li prendre son plume , li assisé tranquille ,
            et li fini écrire sa billlet là moi porté vous.
            Ah vlà li.
        </p>
    </sp>
    <sp who="STRAFFORD" xml:lang="fra-ang">
        <speaker>STRAFFORD lit le billet haut.</speaker>
        <p>» Vous avez raison , monsieur ,
            je suis mort pour vous et pour votre ami » .
            <stage> ( Il parle. )</stage>
            Toi voir lui mort  [etc...]
        </p>
    </sp>
    <sp who="BELTON" xml:lang="met-fr">
        <speaker>BELTON.</speaker>
        <p>Moins que jamais ;
            c'est absolument une énigme pour moi.</p>
    </sp>
</div>
```

Figure 1: This excerpt from Scene 10 of *Le duel singulier* (Dorvigny, 1800) shows how we tag language usage by speaker. It includes standard French alongside Anglo-Baragouin and an unspecified Creole with Mauritian characteristics. [formatting adjusted]

used two complementary rule-based strategies. For plays where one character (or more) uses non-standard speech throughout, we simply identified the ‹sp› (speech) tags associated with that character and inserted an `xml:lang` attribute with the corresponding label, which allowed us to keep associations between characters and speech turns. Additionally, we added tags at the ‹p› level to facilitate text extraction.

For prose, keeping track of specific characters was more difficult. Initially, we tried implementing key-ngram-based regex patterns. Because our languages of interest are frequently embedded in longer French passages, a preprocessing step of sentence tokenization was implemented. Although our disjunctive n-grams generally correspond to words, we use character-level regex patterns that incorporate a special boundary symbol to minimize multi-level tokenization. For the initial annotation, the presence of any one disjunctive n-gram was sufficient to trigger the relevant label. While this method was very useful for highlighting interesting passages, manual retouching was necessary to fix issues of imperfect sentence tokenization, as well as missed examples. In Figure 2, we find a reported clause in Louisiana Creole that is not marked because it contains no disjunctive words, followed by a reporting clause in French [14], that is unintentionally included with correctly identified

[14] dit l'esclave d'une voix caressante' [said the slave with an affectionate voice].

Target/Region	Label	Works	Tokens	Timespan
Normative French	met-fr	35	37066	1649-1779
Peasant	fra-dia	14	27825	1665-1740
Gascon	fra-gsc	4	4530	1672-1800
Anglophone	fra-ang	4	4441	1509-1800
Continental Germanic	fra-deu	25	6899	1580s∼1779
Flemish (Tourcoing/Lille)	fra-nld	4	2664	1880-1932
Réunion	rcf	3	10713	1760s, 1830s
Lesser Antilles (Martinique)	gcf	2	477	1671
Haiti	hat	4	7395	1730s∼1802
Louisiana	lou	10	26068	1748-1895
French Guiana	gcr	2	43414	1796, 1885
Mauritius (tentative)	mau	1	196	1800

Table 2: An overview of the linguistic and temporal spread of the corpus.

Creole speech in the following sentence. The third sentence is marked as expected.

```
<p>
    « Comme vous bel !
    <s xml:lang="lou"> dit l'esclave d'une voix caressante ;
    vou gagnin ain ti lair si tan comifo ! </s>
    <s xml:lang="lou">vou popa riche, mo sûr ;
    di li achté moin.</s>
    ...
</p>
```

Figure 2: Uncorrected semi-automatic annotation of *L'Habitation Saint-Ybars* (Mercier, 1881)

4.4 Compiling Extracts

After adding language tags at the document level, we created a composite timeline that balances facilitating direct comparison between excerpts with giving some level of contextualization. For plays, we extracted scenes where at least one of the ‹sp› turns contained an `xml:lang` attribute with an appropriate value, as demonstrated by Figure 1. By extracting the entire scene, we include samples of normative French and retain the coherence of the conversation to some extent. For monolingual poems, we included the entire poem, albeit possibly excluding meta-linguistic commentary such as notes. For prose, we implemented a multi-level extra process of first trying to identify broad tags like ‹p› based on the `xml:lang` attribute, and then narrower tags like ‹s› only if they were not already included as part of a broader group. In Figure 2, the overall paragraph would be assumed to be French, so only the lines within the ‹s› tags would be extracted, which is why correcting the linguistic annotation is important.

4.5 Balancing

As exemplified by the Gascon accent, the literary conventions can be summarized using a relatively short list of rules. This means that there is a degree of diminishing returns to adding additional examples once we have a basic understanding of said rules. As such, we did not concern ourselves with attempting to create a statistically balanced corpus. In particular, due to the more labor-intensive nature of (semi)-manual encoding, we deprioritized the Peasant French variety early on because it has already received more careful study, and instead focused on the earliest and latest attestations of Baragouin. This may create the impression that literary Peasant French was primarily a 17th century phenomenon. However, this stereotype remained in use until the 19th century. Along similar lines, we did not include many attestations of Mauritian Creole precisely because a digitally accessible diachronic corpus to the same effect already exists (Fon Sing, 2013; Baker et al., 2007).

5 Corpus Presentation

Overall, we found 301 historical works which demonstrate features relevant for the history of FBCLs. We have selected excerpts from 68 of these works to form the basis of the first version of the corpus. The earliest text is “Le Testament du Gentil Cossoys”, written anonymously around 1509, and the most recent is Jules Watteuw’s “Belle Réponse”, published in 1932. The main corpus consists of a single, publicly available XML file containing bibliographic information for the collection, followed by a body which contains “TEI” tags that regroup the relevant selections from each work and are accompanied by their own brief bibliography section. From this file, one can create customized subcorpora that correspond to specific questions by specifying a date range and the language labels that are to be considered.

At present, the corpus contains a total of 188,866 tokens (whitespace tokenization), excluding meta-

Target/Region	Infinitive	Inflected	TAM	CE	Tokens
Normative	105	**1328**	129	254	37066
Peasant	76	**1006**	129	251	27825
Gascon	14	**131**	16	47	4530
Anglophone	**74**	32	7	5	4441
Continental Germanic	**89**	62	11	13	6899
Industrial Flemish	0	**44**	0	18	2664
Réunion	5	**125**	54	2	10713
Haiti	0	157	**102**	27	7395
Louisiana	10	1086	**944**	129	26068
French Guiana	1	1001	**950**	40	43414

Table 3: Attestations of different forms of 'être'. TAM and CE cover creolized inflection.

data. Because of the historical focus of the text, all of the primary sources are in the public domain, and most are readily consultable online. In these cases, we also retain cached copies with additional bibliographic information. In the cases where quotes have been included from printed secondary sources, we do not include metalinguistic commentary. Table 2 provides a high-level summary of the varieties we distinguish and their relative sizes and time spans.

6 Preliminary Results

Since the main effort of this work has consisted of gathering and grouping multiple non-standardized varieties, proceeding directly to quantitative methods presents special challenges. For the initial demonstration, we provide a few qualitative observations and show how we can support them through relatively simple frequency-based methods, with a particular focus on the relevance of Baragouin [15].

6.1 First Person Pronoun: Mo(è)

During the colonial era, the French pronoun 'moi' [me] had two primary variants : **mwe** and **mwa**. FBCLs can be grouped according to which form of 'moi' became the subject pronoun. The first group, consisting of Haitian and Lesser Antillean Creoles, predominantly uses **mwɛ̃**, which is clearly a nasalized version of **mwe** (Hull, 1979). The second group, comprised of Mauritian, Seychellois, French Guianese, and Louisiana Creoles, uses the form **mo**. This division corresponds to further differences in the pronominal system, with the first group also using case-invariant pronouns and marking possession through postposition, while the second group distinguishes between subject and oblique variants and uses proposed possessive adjectives [16]. Although **mo** is tied to **mwa**, its exact origins are less clear. Furthermore, there is documentation that **mo** was once used by the first group, before being replaced in the 1900s (Hazaël-Massieux, 2008).

Several of our documents shed new light on the relationship between these two variants. Firstly, beyond the canonical **mo**, we also found examples of 'moué','moé', 'moè', 'moë', and 'moa' in 19th century Louisiana alone. In Jobey (1860, p.189), for example, includes '*Moè* té cré bien, *moè* perdu papier la yest'[17], which combines the Caribbean-like **mwe** with the Louisiana-specific definite plural marker **laje** (spelled 'la yest'). By itself, this can be explained by 19th century New Orleans' status as a crossroads of French- and Creole-speaking networks. Secondly, however, we found numerous attestations of **mo** in Flemish Baragouin. For example, the opening line of "Poutche" (Watteuw, 1927) is 'Accoute un fos, **mo** ne pas bête' [18]. The latter may help explain **mo** as one innovation which diffused from Europe alongisde mwe, rather than a parallel innovation.

6.2 Copula: ê(tre)

Additionally, we noticed that Baragouin has a tendency to overuse the infinitive 'être' (to be), rather than either conjugating the verb like French, or omitting the copula as in FBCLs. We began quantifying this variation by measuring the frequency of two basic patterns: the infinitive, and all inflected forms. We further tracked two subsets of inflected forms forms that have been integrated into various FBCLs: (precursors of) TMA markers (**(e)te**), **s(r)e**', **s(r)a** and orthographic variants thereof),

[15] The following section uses broad IPA in bold.

[16] Exceptionally, Réunion uses **mwɛ̃** with case distinctions.

[17] I had really though I lost the papers.

[18] Listen up, I ain't [lit. me not] stupid.

as well fusions involving the pronoun 'ce'. For demonstrative purposes, we set aside the samples for the Lesser Antilles and Mauritius, since they are particularly limited. Unfortunately, we could not take into account the clause-restricted copula **je** due to it being homophonous with the much more frequent third-person plural pronoun and a derived plural marker in Louisiana and French Guiana.

Table 3 demonstrates the results of this experiment. As expected, Normative French, Peasant French, and Gascon-accented French all use a wide variety of inflections. In contrast, the FBCLs the Americas retain specific grammaticalized uses, such that 'être' is rare, while inflected forms largely correspond to either TAM markers or presentatives with 'ce' [19]. Réunion, which is distinguished among FBCLs for retaining French auxiliaries, stands out as transitional. In contrast to both groups, both Anglophone and Continental Baragouin (but notably not later Flemish) generalize use infinitive 'être' more than inflected forms, but do not completely discard the latter.

7 Discussion

7.1 Missing (L)(i)nks

By itself, the generalization of 'être' shows that decreased use of inflection and copula deletion, two traits of FBCLs suggested to indicate pidgin origins by McWhorter (2018), did not *necessarily* develop at the same time nor for the same reason. Beyond this, however, we are able to directly tie one process underlying the generation of Baragouin to one Creole language in particular: Réunion Creole (RC).

As Hull (1993, p.393) observes, the subject pronoun **li**, shared by all FBCLs, is employed by a Swiss German in *Le Bourgeois gentilhomme* in place of 'il'. More specifically, as Damm (1911) remarks, the systematic insertion of this third-person pronoun before verbs, as mentioned in Section 3.2.3, is particularly reminiscent of RC. In early texts demonstrating a transitional variety between French and RC, we find sentences such as '**Moi *i* crois** vrai, bien vrai dans mon cœur n'en a bon Dieu' [20] (Bollée, 2007). In both Baragouin and RC, this preverbal pronoun also fuses with auxiliaries, as in this example from *Les filles errantes* (Regnard, 1690): 'Moi **l'être** un étrangir qui cherchir à logir dans sti ville.' [21] and the Réunionese 'Moi **l'est** bien content voir à vous' [22] (Héry, 1883).

Although the exact function and source of the preverbal marker in Réunionese Creole are both debated, one common interpretation is that it marks finiteness on verbs and originated as a generalization of third person reprise pronouns (Bollée, 2007). Interestingly, a similar generalization of third-person **object** pronouns is observed in Spanish-language representations of Africans as early as the 17th century, and comparable phenomena continue to exist in varieties of Spanish in the Americas influenced by Quechua and Nahuatl (Lipski, 2001). In our corpus, we also observe that 'li' in particular also appears in Peasant French, primarily as a clitic indirect object. As Baragouin also inserts preverbal pronouns in sentences that use the French first-person subject clitic 'je', the inserted preverbal pronoun corresponds to a few homophonous French subject, object, and adverbial pronouns. This in turn suggests our data is relevant for contact scenarios beyond FBCLs.

7.2 The Bigger Picture

Beyond tracking individual features, our corpus offers a window into the broader sociolinguistic context of French in the early modern period. In the case of the first person pronoun, despite the temporal mismatch, the specificity of 'mo' points to the Low Countries as a point of interest. Upon closer examination, several works spell out a network connecting Swiss soldiers to this region and Paris in the context of the French-Hapsburg wars such as a 1692 "Air suisse ou flamand" which references the Nine Years' War in Mons, Namur and Maastricht directly, This detail is of interest for Louisiana and Mauritius, where German-speaking settlers and soldiers played important roles in the French colonization in the 1720s. (Vaughan, 2005; Klingler, 2003; Baker and Corne, 1982).

Along similar lines, *Le duel singulier* stands out as a ready-made case study. This play combines normative French, the Gascon accent, Anglo-Baragouin, and an unspecified Creole, as exemplified in Figure 1. As such, it bolsters theories that the FBCLs of the Caribbean region may have developed during the period of Anglo-French cooperation during the early 17th century on islands such as Saint-Christophe and Tortuga (Parkvall, 1995). Fur-

[19] And **je** which we left out as explained above.

[20] I [lit. me **it**] believe true, very true in my heart there is [good] God.

[21] I am [lit. me **it be**] a foreigner looking to lodge in this town

[22] I am [lit. me **it's**] glad to see you.

thermore, the Baragouin can be cross-referenced against the Law French of English courts of that period (Löfstedt, 2014).

8 Conclusion

In short, we have introduced the Molyé corpus, a new resource which puts French literary stereotypes alongside early forms of several French-based Creole languages. We have shown that restructuring of the French pronominal and verbal systems are attested throughout the 16th, 17th, and 18th centuries, and specifically associated with speakers of Germanic languages. Although stereotypes like the conventionalized Baragouin only address a fraction of the real linguistic variation of the period, our corpus nevertheless raises important questions about how people communicated in lands where French and Germanic languages came into contact. Furthermore, it shows that at least some of the divergences between FBCLs and French can be traced back to developments which were already underway in Europe.

Limitations

The major constraint of this work has been converting unstructured works into XML-TEI. As mentioned in the methodology, this involved complete re-transcription in some cases. Overall, we found more than 200 pertinent documents, but were only able to include one third of them. In particular, we had to leave out works in regional languages of France such as Picard, Walloon, and Poitevin. Similarly, we did not address some relevant phenomena, such as the 17th century Carib Baragouin and the 19th century Tirailleur French in order to maintain the scope of the work. Although we are well aware of such varieties, we found few instances using our method, and thus leave them as natural targets for future work.

Ethics Statement

The main idea of this article is that European literary stereotypes from before and during the colonial period can help fill in the some gaps in the early history of (French-based) Creole languages. As such, many of the primary and secondary sources that we have compiled contain negative imagery and commentary regarding various social groups. Sharing such sources should not be taken as endorsement of the views contained therein.

Acknowledgements

This work was primarily funded by the Inria "Défi"-type project COLaF. This work was also partly funded by the last author's chair in the PRAIRIE institute funded by the French national agency ANR as part of the "Investissements d'avenir" programme under the reference ANR-19-P3IA-0001.

References

Wendy Ayres-Bennett. 2000. Voices From The Past: Sources of Seventeenth-Century Spoken French. *Romanische Forschungen*, 112(3):323–348.

Wendy Ayres-Bennett. 2004. *Sociolinguistic Variation in Seventeenth-Century France: Methodology and Case Studies*. Cambridge University Press.

Philip Baker and Chris Corne. 1982. *Isle de France Creole: Affinities and Origins*. Karoma, Ann Arbor, Mich.

Philip Baker, Guillaume Fon Sing, and Vinesh Y. Hookoomsing. 2007. The corpus of Mauritian Creole texts. *The making of Mauritian Creole. Analyses diachroniques à partir des textes anciens*, (9):1–61.

Peter Bakker, Aymeric Daval-Markussen, Mikael Parkvall, and Ingo Plag. 2011. Creoles are typologically distinct from non-creoles. *Journal of Pidgin and Creole Languages*, 26(1):5–42.

Helena Bermudez Sabel, Francesca Dell'Oro, Cyrielle Montrichard, and Corinne Rossari. 2022. Setting Up Bilingual Comparable Corpora with Non-Contemporary Languages. In *Proceedings of the BUCC Workshop within LREC 2022*, pages 56–60, Marseille, France. European Language Resources Association.

Gabriel Bernier-Colborne, Cyril Goutte, and Serge Leger. 2023. Dialect and Variant Identification as a Multi-Label Classification Task: A Proposal Based on Near-Duplicate Analysis. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 142–151, Dubrovnik, Croatia. Association for Computational Linguistics.

Annegret Bollée. 2007. Deux textes religieux de Bourbon du 18e siècle et l'histoire du créole réunionnais.

Isaac Caswell, Theresa Breiner, Daan van Esch, and Ankur Bapna. 2020. Language ID in the Wild: Unexpected Challenges on the Path to a Thousand-Language Web Text Corpus. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6588–6608, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Robert Chaudenson. 1981. Textes créoles anciens: La Réunion et île Maurice: Comparaison et essai d'analyse. *Kreolische Bibliothek*.

Robert Chaudenson. 2001. *Creolization of Language and Culture*. Routledge.

Otto Damm. 1911. *Der deutsch-französische Jargon in der schönen französischen Literatur*. Emli Eberling.

Silvio Moreira de Sousa, Johannes Mücke, and Philipp Krämer. 2019. A History of Creole Studies. In *Oxford Research Encyclopedia of Linguistics*.

Louis Francois Dorvigny. 1800. *Le duel singulier, comédie en un acte et en prose*.

Manuel Favaro, Elisa Guadagnini, Eva Sassolini, Marco Biffi, and Simonetta Montemagni. 2022. Towards the Creation of a Diachronic Corpus for Italian: A Case Study on the GDLI Quotations. In *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*, pages 94–100, Marseille, France. European Language Resources Association.

Charles A. Ferguson. 1975. Toward a Characterization of English Foreigner Talk. *Anthropological Linguistics*, 17(1):1–14.

Charles A. Ferguson. 1981. 'Foreigner Talk' as the Name of a Simplified Register. *International Journal of the Sociology of Language*, 1981(28):9–18.

Paul Fièvre. 2007. Théâtre classique.

Guillaume Fon Sing. 2013. Corpus de textes anciens en créole mauricien.

Simon Gabay, Thibault Clérice, Pauline Jacsont, Elina Leblanc, Marie Jeannot-Tirole, Sonia Solfrini, Sophie Dolto, Floriane Goy, Carmen Carrasco Luján, Maddalena Zaglio, Myriam Perregaux, Juliette Janes, Benoît Sagot, Rachel Bawden, Rasul Dent, Oriane Nédey, and Alix Chagué. 2024. Reconnaissance des écritures dans les imprimés. In *Humanistica 2024*, OCR, Meknès, Morocco. Association francophone des humanités numériques.

Thomas-Simon Guelette. 1740. *Première Parade*. Paul Fièvre.

Walter Haas. 2015. « Déguisé en Suisse » : les « Suisses » de Molière et leur langage. *Littératures classiques*, 87(2):191–189.

Marie-Christine Hazaël-Massieux. 2008. *Textes anciens en créole français de la Caraïbe: Histoire et analyse*. Editions Publibook.

Marie-Christine Hazaël-Massieux. 2013a. Creolica: Revue du Groupe Européen de Recherches en Langues Créoles.

Marie-Christine Hazaël-Massieux. 2013b. Groupe Européen de Recherches en Langues Créoles.

Louis Héry. 1883. *Fables Créoles et Explorations Dans l'intérieur de l'île Bourbon: Esquisses Africaines*. J. Rigal.

Alexander Hull. 1979. On the origin and chronology of the French-based creoles. *Readings in creole studies*, 2:201.

Alexander Hull. 1993. The transmission of Creole languages. *Atlantic Meets Pacific–A Global View Of Pidginization and Creolization (Selected Papers from the Society for Pidgins and Creole Linguistics)*, pages 391–397.

Tommi Jauhiainen, Marco Lui, Marcos Zampieri, Timothy Baldwin, and Krister Lindén. 2019. Automatic Language Identification in Texts: A Survey. *Journal of Artificial Intelligence Research*, 65.

Charles Jobey. 1860. *L'amour d'un Nègre*. Michel Lévy frères.

Amr Keleg and Walid Magdy. 2023. Arabic Dialect Identification under Scrutiny: Limitations of Single-label Classification. In *Proceedings of ArabicNLP 2023*, pages 385–398, Singapore (Hybrid). Association for Computational Linguistics.

Benjamin Kiessling, Robin Tissot, Peter Stokes, and Daniel Stökl Ben Ezra. 2019. eScriptorium: An Open Source Platform for Historical Document Analysis. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, volume 2, pages 19–19.

Alain Kihm. 2018. Língua de Preto, the language of the African slave community in Portugal (16th–19th centuries). *Language Ecology*, 2(1-2):77–90.

Ruth King, Terry Nadasdi, and Gary R. Butler. 2004. First-person plural in Prince Edward Island Acadian French: The fate of the vernacular variant je… ons. *Language Variation and Change*, 16(3):237–255.

Thomas Klingler. 2003. *If I could turn my tongue like that: the Creole Language of Pointe Coupee Parish, Louisiana*. LSU Press.

Jacques Landrecies. 2001. Une configuration inédite : la triangulaire français-flamand-picard à Roubaix au début du XXe siècle. *Langage et société*, 97(3):27–69.

Elizabeth Eva Leach. 2020. Ripping Romance to Ribbons: The French of a German Knight inThe Tournament at Chauvency. *Medium Ævum*, 89(2):327–349.

Heather Lent, Emanuele Bugliarello, and Anders Søgaard. 2022a. Ancestor-to-Creole Transfer is Not a Walk in the Park.

Heather Lent, Kelechi Ogueji, Miryam de Lhoneux, Orevaoghene Ahia, and Anders Søgaard. 2022b. What a Creole Wants, What a Creole Needs. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6439–6449, Marseille, France. European Language Resources Association.

John M. Lipski. 2001. Bozal Spanish: Restructuring or creolization? *Degrees of Restructuring in Creole Languages*, 22:437.

Johann-Mattis List. 2024. Open Problems in Computational Historical Linguistics. *Open Research Europe*, 3:201.

Nikola Ljubesic, Nives Mikelic, and Damir Boras. 2007. Language Indentification: How to Distinguish Similar Languages? In *2007 29th International Conference on Information Technology Interfaces*, pages 541–546.

Anthony Lodge. 1991. Molière's Peasants and the Norms of Spoken French. *Neuphilologische Mitteilungen*, 92(4):485–499.

Anthony Lodge. 2003. *French: From Dialect to Standard*. Routledge, London.

Leena Löfstedt. 2014. Notes on the beginnings of Law French. *Romance Philology*, 68(2):285–337.

Pierre Carlet de Chamblain de Marivaux. 1727. *L'île de La Raison*. Paul Fièvre.

Pierre Carlet de Chamblain de Marivaux. 1735. *La Mère Confidente*. Paul Fièvre.

Oliver Mayeux. 2019. Rethinking decreolization: Language contact and change in Louisiana Creole.

John H. McWhorter. 1998. Identifying the Creole Prototype: Vindicating a Typological Class. *Language*, 74(4):788–818.

John H. McWhorter. 2018. *The Creole Debate*. Cambridge University Press.

Alfred Mercier. 1881. *L'habitation Saint-Ybars: Ou, Maitres et Esclaves En Louisiane, Rećit Social*. Nouvelle-Orléans. Imprimerie franco-américaine (E. Antoine.

Ingrid Neumann-Holzschuh. 1987. Textes anciens en creole louisianais: Avec introd., notes, remarques sur la langue et glossaire.

Mikael Parkvall. 1995. The role of St. Kitts in a new scenario of French Creole genesis. *From Contact to Creole and Beyond*.

Jean-François Regnard. 1690. *Les filles errantes*. Paul Fièvre.

Nathaniel Robinson, Raj Dabre, Ammon Shurtz, Rasul Dent, Onenamiyi Onesi, Claire Monroc, Loïc Grobol, Hasan Muhammad, Ashi Garg, and Naome Etori. 2024. Kreyòl-MT: Building MT for Latin American, Caribbean and Colonial African Creole Languages. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3083–3110.

Nathaniel Romney Robinson, Matthew Dean Stutzman, Stephen D. Richardson, and David R. Mortensen. 2023. African Substrates Rather Than European Lexifiers to Augment African-diaspora Creole Translation. In *4th Workshop on African Natural Language Processing*.

Pablo Ruiz Fabo, Delphine Bernhard, and Carole Werner. 2020. Création d'un corpus FAIR de théâtre en alsacien et normalisation de variétés non-contemporaines. In *2èmes Journées Scientifiques Du Groupement de Recherche Linguistique Informatique Formelle et de Terrain (LIFT)*, pages 34–43, Montrouge, France. CNRS.

Patrick Sauzet and Guylaine Brun-Trigaud. 2015. La Lucette de Monsieur de Pourceaugnac : " Feinte gasconne ", vrai occitan. *Littératures classiques*.

Charles-Timoléon de Beauxoncles de Sigogne. 1620. Testament d'vn Escossois. In *Le Cabinet Satyrique...* Billaine.

David Baird Smith. 1920. Le Testament du Gentil Cossoys. *The Scottish Historical Review*, 17(67):190–198.

Anand Syea. 2017. *French Creoles: A Comprehensive and Comparative Grammar*. Routledge.

TEI Consortium eds. 2023. TEI P5: Guidelines for Electronic Text Encoding and Interchange.

Sarah Grey Thomason and Terrence Kaufman. 1988. *Language Contact, Creolization, and Genetic Linguistics*. University of California Press.

Megan Vaughan. 2005. *Creating the Creole Island: Slavery in Eighteenth-Century Mauritius*. Duke University Press, Durham, NC.

Jules Watteuw. 1927. Pitche.

Henri Wittmann. 1995. Grammaire comparée des variétés coloniales du français populaire de Paris du 17e siècle et origines du français québécois. *Le français des Amériques, dir. Robert Fournier & Henri Wittmann*, pages 281–334.

Jonathan Wylie. 1995. The Origins of Lesser Antillean French Creole: Some Literary and Lexical Evidence. *Journal of Pidgin and Creole Languages*, 10(1):77–126.