WaCadie: Towards a Web Corpus of Acadian French

by

Jérémy Robichaud

Bachelor of Computer Science, UNB, 2022

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

Master of Computer Science

In the Graduate Academic Unit of Computer Science

Supervisor(s):	Paul Cook, PhD, Computer Science
Examining Board:	Michael Fleming, PhD, Computer Science, Chair
	Przemyslaw Pochec, PhD, Computer Science
	Christine Horne, PhD, French

This thesis is accepted by the
Dean of Graduate Studies

THE UNIVERSITY OF NEW BRUNSWICK

December, 2023

Abstract

Corpora are important assets within the natural language processing and linguistics communities. However, not all low-resource languages have corpus representation. Acadians, an eastern people of North America, do not have a corpus representation of their variation of French. An Acadian French corpus could allow for a better understanding of the unique dialect. Leveraging web-as-corpus methodologies such as BootCaT, domain crawling, and social media scraping, we create three different corpus representations of Acadian French. Each corpus is, on its own, an Acadian French resource while also showcasing the strengths of their individual method of creation. We propose 22 statistical corpus-based measures stemming from previously researched Acadian French characteristics to compare these newly built corpora to known Acadian French text. We found that while all three yield traces of Acadian French text, BootCaT is the largest corpus, and social media scraping has the highest count of Acadian French characteristics.

Acknowledgements

I want to acknowledge all my friends and family who helped me through this journey. I could not have completed it without you all. I would also like to acknowledge Dr. Paul Cook, my supervisor, for supporting, teaching, encouraging, and mentoring me throughout this process. I would like to thank him for pushing and working alongside me to achieve goals that I did not think possible.

Table of Contents

List of Tables

Chapter 1

Introduction

A corpus is a collection of texts that can be in one or more languages, annotated, and digitalized [24]. Corpora are significant assets within the natural language processing (NLP) and linguistics communities as they allow the training of models and corpus-based study of languages [41, 29, 42, 32]. Due to their high utility, many different methodologies have been developed to create corpora. However, gathering the appropriate text is one of the biggest challenges in creating a corpus. With 4.70 billion users in 2020, the internet has become a large information bank.[1] This large information bank can be leveraged to create corpora in multiple ways [58]. This corpus-building process is known as web-as-corpus (WaC) and can be used to create corpora of many different languages.

Since the internet is so large and diverse, various ways of creating a web corpus exist. In 2004, researchers used search engines (e.g., Google) to query URLs based on words related to the targeted corpus [9]. This method became known as BootCaT. Another method is leveraging the organization of websites on the internet. Each website has a specific link (e.g., `https://www.unb.ca/`), called Uniform Resource Locator (URL), where each URL section (e.g., *https*, *www*, *unb*, *ca*) has a unique meaning. The last portion (e.g., *.ca*) is called the domain, and it represents a certain group to which

[1] `https://ourworldindata.org/grapher/number-of-internet-users`

URLs belong (e.g., *.ca* means the Canada domain). We can use these URL structures to gather websites within a certain group and extract texts from them [21]. This is called domain crawling. In recent years, services have appeared which offer large databases of downloaded websites. An example of this service is Common Crawl [67]. This pairs well with domain crawling, allowing easy access to websites with a specific domain [26]. Lastly, social media has become immensely popular, reaching 4.59 billion users in 2022.[2] Every day, new posts are created by users. These posts hold information, including text. These posts are now being extracted and used to create corpora [8]. This has become known as social media scraping.

An important value of creating corpora is preserving languages, especially for under-represented languages, such as minority variants of languages. Minority variants of languages are languages in smaller communities that have non-standard features [57]. Regional dialects from two locations show this since two locations could speak the same language but have different spellings for the same word. An example is *wrecked* and *banjaxed*. *Wrecked* is a word known throughout many English-speaking regions, whereas *banjaxed* is a regional alternative seen by English speakers in Ireland [49]. This regional language, Hiberno-English, is considered a minority variant of the English language. As such, obtaining corpora of minority variants of languages is useful for obtaining a more complete view of the nuances of the language. BootCaT, domain crawling, and social media scraping were all previously used to create corpora of minority variants of languages, such as national varieties of English [21] and Egyptian Arabic [3].

In the early 17th century, a group of French colonizers arrived on the shores of what is now Nova Scotia and Prince Edward Island and named the land *Acadia* [31]. The people of Acadia shortly after that became known as *Acadians*. In the 18th century, the Acadians were forced out of their homes and scattered across the eastern parts

[2] https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/

of North America [31]. Today, Acadians can still be found in a wide variety of eastern regions of North America. Most Acadians reside in the Atlantic Provinces of Canada. However, the largest portion resides in New Brunswick [31, 6, 68]. Acadians can also be found in the regions of Maine and Louisiana of the United States of America, where they can also be known as *cajun* [33]. Since they first began as French explorers, most Acadians speak French; however, throughout the years, the French spoken by Acadians changed into a unique variety of French that we now call *Acadian French* [7]. Additionally, multiple varieties of Acadian French exist, partially due to the influence of the surrounding colonies [7]. Examples of these varieties are *Chiac*, which can be found within the south-east region of New Brunswick [11] and *Brayon* in the North-West region of New Brunswick [4].

To our knowledge, little research has been done on the presence of Acadian French on the web. There currently does not exist a large corpus of Acadian French, but there exist a few smaller spoken corpora [68, 5, 11, 50].

We begin our research by asking *How well do NLP tools perform on Acadian French text?* Evaluating the performance of current-day NLP tools on Acadian French could indicate the need for an Acadian French corpus. Previous work showed that these tools could perform well on French [44], but whether this applies to a minority variant of French, specifically Acadian French, is unknown. We will assess if these tools perform well on Acadian French in Chapter 3 by comparing the performance of two modern-day NLP tools on standard French and Acadian French. For this evaluation, we built an Acadian French corpus using an Acadian French dictionary, which can be found in Chapter 3.1. The first test is how well a language model predicts words in both texts (Chapter 3.2). The second test compares a part-of-speech tagger on the corpora (Chapter 3.3). Both tests suggest that Acadian French is harder to process than standard French for these tools. A potential solution could be creating a large Acadian French corpus to help train the tools and adapt them

to Acadian French. This thesis examines potential solutions for creating a large Acadian French corpus by asking two additional questions.

We follow by asking *Can we create an Acadian French corpus using previously proposed web-as-corpus methodologies?* Before answering this question, we need to create a method of identifying Acadian French text to confirm if the texts acquired from the web are truly Acadian French. Leveraging knowledge of Acadian French linguistic characteristics from numerous sources, we propose 22 corpus-based statistical measures to help identify Acadian French text (Chapter 5.3). To create a point of comparison, our study uses the Acadian French dictionary [22] as a source of known Acadian French. We created a corpus from the example sentences found in each dictionary entry. In contrast, we created a similar corpus from Wiktionary, an online French dictionary.[3] This gave us two comparable corpora, both consisting of French dictionary definitions, one Acadian and one not, to confirm if our measures are indeed indicators of Acadian French. Finally, we use these measures on three newly built corpora from different previously proposed web-as-corpus methodologies. Our results (Chapter 5.5) demonstrate that corpora built from BootCaT, domain crawling, and social media scraping all show signs of Acadian French.

Lastly, we ask *Is there one of these newly built corpora that surpasses others in quality and quantity?* We first create a smaller corpus from known Acadian French text (Chapter 3.1) to create an Acadian French reference corpus. Then, we compare all three of our newly built web corpora and the reference corpus to a known standard French web corpus (Chapter 5.4). While all three newly built web corpora showed signs of Acadian French, the BootCaT and social media scraping corpora showed similar or greater amounts of Acadian French characteristics than our Acadian French reference corpus. Our social media scraped corpus had the highest number of Acadian French characteristics but also had the smallest size. BootCaT had a similar

[3] `https://fr.wiktionary.org/`

number of Acadian French properties to our benchmark Acadian French corpus while outsizing our benchmark corpus by two orders of magnitudes in terms of number of tokens. A token is a set of characters in a document corresponding to a word or punctuation. The social media scraping corpus is sized at 50k tokens, the domain crawling corpus at 377k tokens, and the BootCaT corpus at 1.2M tokens.

This thesis is structured as follows. Chapter 2 presents past research and background information on traditional corpus creation (Chapter 2.1), web corpus creation (Chapter 2.2), web corpus creation for minority languages (Chapter 2.3) and Acadian French (Chapter 2.4). Chapter 3 shows how off-the-shelf NLP tools perform on Acadian French by creating reference corpora (Chapter 3.1) and using them to compare the performance of a masked word prediction model (Chapter 3.2) and a part-of-speech tagger (Chapter 3.3). Chapter 4 describes the corpora created (Chapter 4.1) and the post-processing pipeline used (Chapter 4.2). Chapter 5 starts by presenting the known French web corpus we will use as a reference corpus (Chapter 5.1), followed by analyzing the keywords of the potential Acadian French corpora created (Chapter 5.2). Then we continue by showcasing the 22 statistical measures we created to identify Acadian French text (Chapter 5.3), assess those measures on known Acadian French text to prove their functionality (Chapter 5.4), and then present the results of using these measures on all our Acadian French web corpora created (Chapter 5.5). Chapter 6 summarizes the work, outlines the conclusions and discusses potential future work stemming from this work.

Chapter 2

Related Works

To provide the reader with a better understanding of this thesis's research and scope of work, the presentation of some background information is required. The first is a brief overview of traditional and contemporary corpora. The second is a broad examination of the previous web-as-corpus work, followed by web-as-corpus work that dealt with minority varieties of languages. Lastly, we will discuss the prior research done with the Acadian French dialect.

2.1 Traditional Corpora

In many computational linguistic studies, a large corpus is needed. Dating back to the 1960s, multiple methods of corpus creation have been explored. This section will explore past work related to traditional corpora, i.e. non-web corpora.

2.1.1 Notable Corpora

One early corpus is the *Standard Corpus of Present-Day American English*, now commonly known as the Brown Corpus [41, 29]. The Brown Corpus consists of written text by American English writers no later than 1961. The corpus holds more than one million tokens. Its size stems from combing 500 two-thousand-word

samples of English literature. The project started in 1964 and organized the text into 15 categories, each with multiple subtypes. Years later, the corpus was updated to hold part-of-speech tags alongside the text data [29]. They also strived to set a standard pattern for building new corpora [29]. Due to its popularity, multiple offsprings were created. One of them was its British English equivalent, Lancaster-Oslo-Bergen (LOB) Corpus [61].

In 1992, a new corpus called the British National Corpus [42] was created. The British National Corpus has 100 million words of British English compiled from two different media, written and spoken. The researchers noted that such a corpus could enable the studies of fields such as linguistic theory, language variation, and language acquisition. The corpus consists of multiple sources of electronic text and annotated speeches. They also delve into three important parts of corpus creation: acquisition, preparation, and distribution. They remark that making the corpus publicly available and easily accessible is one task that must be upheld after completely processing the data.

In 1996, a new international corpus of English varieties (ICE) was created [32]. It stemmed from the popular Brown and LOB corpora and the various studies comparing them. Due to a lack of spoken English in both corpora, the comparisons were restricted to written text [32]. The ICE comprises 18 national or regional corpora, such as Canadian English and Australian English, each created from text samples similar to the Brown and LOB corpora. However, unlike both of these corpora, more than half of the text from those corpora derives from spoken content.

As such, corpora have been an important resource in linguistic studies, especially computational linguistics. These works also lay the foundation for appropriately presenting newly developed corpora. However, these corpora are all English. In early work, corpus creation emphasized English literature. In the years to follow, corpora were created for diverse languages.

2.1.2 French Corpora

Originating in 1990, the Canadian Hansard is one of the most popular English/French parallel corpora [14]. A parallel corpus consists of documents in two or more languages that directly reflect each other in translation. The corpus was created from documents from the Canadian parliament meetings, which are mandated to be recorded in English and French, called Hansards. This gives access to numerous documents with a one-to-one translation to the other language. The original corpus was composed of roughly 100 million tokens of text in English, which was translated to French by a parliament-appointed translator. The Canadian Parliament Hansards are stored publicly; therefore, the corpus can be easily recreated.

In 2001, the European Parliament minutes were published online. This held lengthy text documents in which multiple members spoke in 11 different languages. These records lead to the Europarl corpus, a multilingual corpus holding 20-30 million tokens for each of 11 languages [40]. These collections of languages were originally used for translation tasks in which text was translated using clustering to another language within the Europarl corpus [40]. The Europarl corpus was applied to statistical machine translation applications in which they translated to and from 110 language pairs [40].

Annotated French corpora, when compared to English, are rare. In 2000, a new French Treebank corpus was created [1, 2]. The French Treebank corpus holds one million tokens sourced from newspapers. The French Tree Bank was used for further corpora and linguistic analysis by numerous other researchers [2].

Like Europarl, in 2010, United Nations documents were leveraged as sources for multilingual text to build a corpus named MultiUN [27]. The documents were gathered from January 2000 to September 2009, containing all six official languages of the UN: Arabic, Chinese, English, French, Russian and Spanish. Each language has a subset of documents averaging 300 million tokens per language. The proposed purpose of

this corpus was machine translation.

In recent years, French has had its fair share of content within corpora [14, 40, 27]. These corpora cleverly use publicly available text material. While we are interested in French text, we are specifically interested in a minority variety of French, Acadian French.

2.1.3 French Corpora of Minority Varieties

In 2011, the first pan-francophone corpus on varieties of French in North America was created called FRAN [45, 46]. It was part of a broader project measuring the French varieties in each continent. The purpose of the corpus is to measure the evolution of French varieties through historical written-based texts and modern oral-based texts. It includes documents from all over North America but primarily from Canada. It holds Acadian French, Québec French, Ontario French, West Canadian French, and United States French texts.

Like other past work, a corpus of the Québec French dialect was created by leveraging online public transcripts of the National Assembly of Québec [47]. The corpus spans from the early 1900s to 2022. However, the distribution curve of tokens per year indicates that most text was recorded past 1963. They also noted some limitations within the corpus, like the fact that the first woman elected was in 1977, and even since then, women have had fewer records of debates than men. As such, the corpus includes predominantly male authors. Even though they do not conclude anything regarding the dialect harvested, they indicate that future work could utilize this corpus to fine-tune models of Québec French.

Both these corpora show that Canadian French dialects have been studied in corpus linguistics. Some past research gathered work on varieties of North American French [45, 46], whereas others focused more on a local variant of French [47]. This is a great indication of interest in varieties of French, whether regional or provincial.

Past work showed interest in an Acadian corpus. New Brunswick researchers created a speech corpus of 140 speakers from assorted locations within New Brunswick; they called it *Reconnaissance automatique de l' acadien* (RACAD) [18]. The speakers were equally distributed on sex, age, and occupation. The main difference was their residing location. The speakers read 212 sentences specifically selected to showcase their dialect. While this corpus was created for speech recognition, more phonetic analysis could be done on the recordings as the speakers showed dialectal differences. While Acadian French was the article of this past research, up to now, the largest Acadian French corpus, to our knowledge, is RACAD, which is spoken. In the hopes of studying the written nuances of Acadian French, we aim to create a text-based corpus.

2.2 Web-as-Corpus

As the web grew in popularity, became larger in size, and more accessible to the average user, researchers began exploring the possibilities of utilizing the sea of data as a source for document collection [58]. Since the web holds numerous languages and genres, web-as-corpus has been popular for the last two decades, and numerous data collection methods have been explored [28, 58, 21, 13]. Like traditional corpora, web-as-corpus combines many small texts (web pages) into a larger corpus. The web offered many advantages that previous corpus construction methods lacked: the cost of construction is inexpensive due to the accessibility of the web, and the size could be very large. However, it also offers different challenges. Noise (e.g. links, photos, toolbars, repeated content) is an example of this since it is content that does not add value to the corpus [58]. Another challenge is unregulated text formatting within the data (i.e. different HTML formatting, different encoding, different spellings of words) as sites do not have a set structure [58]. This section explores three methods of web

corpus construction and works showcasing them. Some work may combine multiple methods; however, we categorized them based on the core harvesting method.

2.2.1 BootCat

In 2004, researchers introduced the BootCaT method to gather large amounts of data from the web [9]. BootCaT uses commercial search engines (e.g., Google or Yahoo) to query sites that may hold valuable text. It searches for medium-frequency tuples of words (also known as seed words) expected to be found within the targeted corpus, collects the resulting URLs, and extracts the text. This allows for language and topic-specific searches. However, since search engines are black box tools, BootCaT gives very little control over what URLs are returned. Regardless, it has been proven to be an effective method for acquiring topic-specific text (e.g., psychiatric articles in both English and Italian [9]).

Following suit, another corpus using the BootCaT method was created, this time in Japanese [64]. They built it using triplets of Japanese words and querying Google for them. They examined the diversity of sources retrieved for the corpus, from source (e.g., blogs, comments, reports) to domain (e.g., business, arts, leisure). They concluded that this technique will likely gather great diversity in both source and domain.

Shortly after, researchers further applied this method when they created the WaCky initiative [28]. This initiative created a collection of monolingual web corpora that is composed of a wide variety of languages. The paper introduces three corpora: English (ukWaC), German (deWaC) and Italian (itWaC) [28]; French (frWaC) was later added to the collection as well.[1] These corpora are created using the seed words obtained by dialectal-specific corpora (e.g. the British National Corpus was used for ukWaC). These seed words were then used to retrieve documents with BootCaT.

[1] https://wacky.sslmit.unibo.it/doku.php?id=corpora

These documents were further used as a basis for web crawling to extend the size of the corpora further. Additionally, they restricted the search by domain to incentivize the result to return dialectal-specific URLs.

BootCaT is efficient at harvesting large amounts of topically relevant documents. It does so, not without drawbacks, however. Due to it relying on black-box search engines, we are unsure if the content returned is appropriately sourced, especially when looking for location-based text. Additionally, we cannot be sure that all website results are equal. The black-box search engine may or may not have a bias towards which website it showcases. Another issue is it is almost impossible to reproduce the corpora built with this method since websites are constantly added, deleted, or even changed, affecting the final acquired corpus. On top of this, the search engine could also get updated, which could alter the results. However, the benefits or corpora created with BootCaT easily outweigh their drawbacks as this approach is capable of creating large corpora of targeted languages [9, 64, 28]

2.2.2 Domain Crawling

In 2017, researchers showed evidence of the effectiveness of domain crawling as a method of corpus construction, as it can yield high-volume region-specific corpora [21]. Instead of gathering websites through searches, domain crawling leverages the URL of websites and collects the data of all websites crawled within a specific domain. This method is often utilized for region-specific domains, such as .ca or .uk, as collecting a large sample of these websites often leads to a corpus containing a region-specific variety of a language [21]. The researchers used domain crawling to effectively show that national top-level domains (e.g., *.ca* and *.uk*) yield content that reflects their corresponding national variety of English [21]. However, domain crawling does not control the content or language within the resulting corpus without post-processing.

An issue of domain crawling is that reproducing the corpora is almost impossible with this method, similar to BootCaT, since websites are constantly added, deleted, or even changed, affecting the final acquired corpus. Thankfully, there are ways around this issue. We can use the Common Crawl, a non-profit organization, database to combat this [67]. Common Crawl periodically releases static snapshots of the web, which can then be used to download all HTML files linked to websites with a specific URL domain. This enables easy reproducibility since the snapshots will not be altered while also constantly adding new snapshots that may add new content to the corpora.

Past work demonstrates this concept by building corpora from Common Crawl using multiple country domains [26]. Their corpora range from different countries and languages. They compare their findings to another corpus creation method, social media scraping. They collected a corpus with 423 billion words spanning over 148 languages, each with a minimum of 1 million words, all gathered from data given by Common Crawl. They used country domains as a source of truth as to the website's location of origin.

Domain crawling is useful for gathering location-based content. However, the content is limited by the snapshots of Common Crawl and the websites found and not found within it. This method also relies on the crawled domain and, when using a regional domain, relies on the standards set on that domain by its region. Regardless, it has been shown to be an efficient method for creating corpora of regional varieties [21, 67, 26].

2.2.3 Social Media Crawling

Lastly, informal text and text representing spoken dialects can also be gathered from the web. A source of such text is social media. Past work addresses the difficulty of using social media as a source of content for a corpus, as well as the richness

of its content [8]. They compare corpora built from different user-annotated text sources (e.g., blogs and forums) to Twitter text. They conclude that while social media-built corpora can be noisy, they can be cleaned and that noise is an issue that can be dealt with using NLP tools such as part-of-speech (POS) tagging and removing non-linguistic tokens. They also conclude that while social media text is less grammatically structured than standard edited text, the differences are relatively small.

Researchers showed the possibilities of using social media, specifically Reddit, to create a corpus focusing on sarcasm [37]. They use known social media cues of sarcasm (e.g., */s*) to gather the data they deem appropriate for the corpus. Their corpus consists of 533M tokens and over 1 million sarcastic comments. This shows the capabilities of using social media, specifically Reddit, as a source of nuanced dialogue.

Additionally, researchers used Reddit as a source of German literature in creating a corpus [13]. Their corpus is 270 million tokens, spanning 380,000 posts. They scraped data from multiple German-based subreddits (subgroups of Reddit). This allowed them to curate German text from a variety of topics. They started with 43,193 subreddits but ended with 2,429 after post-processing. They pruned subreddits based on the percentage of posts and comments in German and only took those above a threshold. This pruning made for a higher concentration of German content left within the corpus.

Social media crawling is useful for gathering user-generated content. This is especially helpful for building corpora that include nuances of speech, such as sarcasm [37], as well as targeted languages [37].

As seen, multiple ways exist of gathering valuable web data to create a corpus. As previously seen, all three methods previously discussed can be used and have been used to create web corpora of a specifically targeted language [9, 64, 28, 21, 67, 26,

8, 37, 13]. Each of these approaches is unique and has its advantages. As such, it is difficult to categorize one method as better than the others. While our research aims to create a corpus of Acadian French, all three methods are potential ways to create such a corpus.

2.3 Web-as-Corpus for Minority Variants of Languages

A variety of a language is a language spoken by a specific group, such as geographically or religiously based, that uses the language in a different, non-standard way [57]. A minority variant is a language variation in which a minority speaks it [57]. To my knowledge, no previous research has created a large corpus of Acadian French, a minority variety of French. Fortunately, previous work showcased the effectiveness of web-as-corpus when creating a corpus of a minority language variant. In this section, we separate the works previously done into three sections, reflecting the three methods of creating web corpora previously considered.

2.3.1 BootCaT

Researchers have shown that seed words and search engines can be used to create a Norwegian web corpus [34]. They referred to this corpus as noWaC. noWaC used domain filtering alongside the BootCaT method to only include *.no* domain websites. It holds 690 million tokens, from 85 thousand URLs remaining after post-processing. They mention the challenge of building a corpus for languages with a small internet presence. However, following this approach gives a corpus roughly one-third the size of the published WaCky corpora.

Past work showed that minority variants of widely-spoken languages could benefit from web-as-corpus methodology by successfully creating a Hiberno-English corpus

using the previously mentioned BootCaT approach [49]. They made a set of 3000 3-tuple and 3000 2-tuple search queries using 81 Irish English seed words, i.e. words specifically belonging to Irish English. They gathered 10 URLs per tuple searched using Yahoo's search engine. The results showed that the newly built corpus exhibited many properties of Irish English.

Researchers created a worldwide English corpus [23]. They queried URLs using Google and filtered them to remove non-regional websites using Google's advanced search features. The corpus spans 20 countries, 1.79 million web pages, and 1.9 billion tokens. They then analytically compared English subsets and analyzed their syntax, morphology, and differences in meaning. This was similar to the ICE corpus [32]. However, it was done on a larger scale.

As we can see, the BootCaT approach yields large corpora of minority varieties. However, this is often coupled with some filtering (often domain filtering) to ensure the corpus is also of high quality. This could be beneficial in creating an Acadian French corpus, as scouring the web for any trace of Acadian French would hopefully yield a decent-sized corpus.

2.3.2 Domain Crawling

Similarly, past research used domain crawling to create corpora of multiple national varieties of English [21]. They used ClueWeb09 [16], a 2009 snapshot of crawled websites filtered by domain and language, to create corpora of national varieties of English.

Researchers created a corpus of global language representations spanning over 140 languages and 150 countries, totalling 423 billion tokens [26]. Their corpus was created using Common Crawl.[2] They compare their newly created corpus to a worldwide Twitter corpus and population demographics to analyze the amount of each

[2] `https://commoncrawl.org/`

language found within the corpus. Their observation was that there seems to be a correlation between the amount of text in a language and the following: the country's population, access to the Internet, and per capita GDP.

Domain crawling is a great method of building a corpus of minority language variants. It yields large and high-quality corpora while confirming the location of the sourced text.

2.3.3 Social Media Crawling

In 2012, researchers created an Egyptian Arabic corpus using Twitter-based querying [3]. They searched known Egyptian Arabic exclusive terms using the Twitter API over seven months, totalling 15M search queries. They combined this with online forums (e.g. Q/A sites) and blog-based search engines to create an Egyptian Arabic corpus of 11M words.

Later, it was shown that it is possible to create a Twitter-based corpus of Brazilian Portuguese [15]. They used a Python Twitter wrapper to harvest 15,000 tweets. These tweets were selected based on whether they had hashtags about popular Brazilian TV shows. Annotators then categorize these tweets as positive, negative, or neutral. They used these classifications for a polarity classification task. However, they also note the relevance of their corpus for linguistic analysis as well.

Researchers used Reddit as a source of German literature in creating a corpus [13]. Their corpus is sized at 270 million tokens, spanning from 380,000 posts. They scraped data from multiple German-based subreddits (subgroups of Reddit), such as r/germany, r/Switzerland, and r/fcbayern. They created the corpus for linguistic analysis. They also note that subreddits have different levels of dialect presence. As such, the heavily dialectal-influenced subreddits should likely be treated separately. Social media corpora are a great source of dialectal information. Acadian French, a dialect of French, might benefit from a corpus from one of these platforms.

All three cases demonstrate that the methodology discussed in Section 2.2 functions for minority variants of widely-spoken languages. This indicates that it may apply to other minority variants as well, such as Acadian French. This research will use these methods to create new corpora of Acadian French.

2.4 Acadian French

Acadians are a linguistic minority in the Atlantic region of Canada, with the largest population residing in New Brunswick [6, 68]. New Brunswick has different variants of Acadian French. For example, Chiac is the most spoken version of Acadian French in the South-East, while in the North-West region, Brayon would be heard more often. Brayon resembles Quebec's French [4], whereas Chiac is a variety of French containing many English borrowings [11]. While Brayon and Chiac are different, both are commonly used Acadian variants and, as such, are targeted equally in our research. Previous linguistic work has been done describing the nuances of Acadian French. Previous work compared Acadian French and Chiac [52, 53]. Another previously done study considered the history of Acadian French and how it evolved into Chiac while noting patterns of speech found during interviews with Chiac speakers [11]. Similarly, other work explored the identity of Brayon, giving historical context and linguistic comparison between it and standardized French [4].

Previous researchers studied spoken Acadian French by creating a spoken corpus [18]. In their research, they emphasized the different properties of Acadian French in each region: North-East, North, and South-East. They note that these regions could vary in certain phonetic pronunciations, such as Brayon more often pronounce *moi* or *toi* as *mwa* or *twa*. However, they also note that general Acadian features exist (e.g. *diable* pronounced *djab* and *guerre* pronounced *djuerre*), which could be seen in all three regions.

Numerous studies dive into understanding the nuances of Acadian French [18, 11, 52, 53, 4]. This is important to our work as it reveals ways of indicating that text may or may not be of the dialect. Understanding the nuances of the dialect helps indicate the quality of text harvested and if it reflects the dialect.

2.4.1 Brayon

In 1992, researchers evaluated the usage of the diphthong *oi*, pronounce [wa] (e.g. m*oi*, t*oi*), in Brayon French [36]. Their work was conducted on a sample of 21 Brayon speakers. The researchers interviewed the Brayon speakers and noted their verbal responses. They concluded that the pronunciation of older Brayon speakers resembled more traditional Acadian French pronunciations of *oi*, pronounced [we], whereas younger Brayon speakers had a tendency to pronounce it [wo]. However, they also remark that [wo] approaches the standard pronunciation of [was]. They concluded this was perhaps because speakers in neighbouring Québec who pronounce it [wa] may have influenced them.

In 2021, research was conducted on the linguistic identity of Brayon [4]. Specifically, they focus on the *Brayonnaire* [59, 60], a locally made dictionary composed of roughly 1000 words and expressions. They explore the phonetic and syntactic nuances of the expressions and words written and the linguistic identity of Brayon. They demonstrate that Brayon French has unique properties only exhibited within their variety of French, such as a unique vocabulary (e.g. *picasser* ('to play with your food'), *pètantoune* ('fat')). They also observe that the phonological structure of their words is borrowed mainly from a combination of Québec and Acadian French.

Brayon French, while still being a subset of Acadian French, has unique features not seen throughout all of New Brunswick [59, 60]. Identifying Brayon French text specifically would also indicate Acadian French text. Something important to consider is how Brayon differs not only from Acadian French but also Québec French. Québec

French, having influenced the evolution of Brayon, could be accidentally viewed as Brayon French due to their overlap. Understanding the differences between the two varieties is crucial to identifying Brayon French.

2.4.2 Chiac

In 1995, research was conducted on how Chiac French speakers can reconstruct sentences since the dialect includes English and French words [51]. They built a small oral corpus of interviews with Université de Moncton students. They did so to evaluate how English imprinted on Acadian French to create Chiac. They observed borrowings from English were present in Chiac French, such as English words (e.g. *love affaires*, *about*, *movie*) and English verbs "transformed" to French (e.g. English *worry* to Chiac *worrier*).

In 2014, researchers analyzed Chiac French borrowings from other languages [52]. They used a small conversation-based corpus for their analysis. They not only compare to other language variants but also to other spoken corpora from past research. They concluded that Chiac borrows most from English (e.g. *because/'cause*, *although*, *too much*), so much so that it distances itself from traditional Acadian French.

In 2020, Chiac French was further analyzed, exploring the people's history, the language's syntax, and the language's meaning for its speakers [11]. They interviewed a group of Chiac speakers across the Greater Moncton region. As their main focus was linguistic identity, they noted that Chiac speakers differ in language identity as most were identified as having linguistic insecurities. The Chiac speakers noted that Chiac was sometimes seen as "moins bon" ('less good') than traditional French.

As seen, there has been a great interest in Acadian French. Yet, all corpora built in the past were primarily based on spoken interviews and not written content. As new corpus construction methods have developed to facilitate corpus creation, perhaps Acadian French could benefit from a large written corpus. Additionally, no measure

has been created to calculate how much a corpus corresponds to Acadian French. Our research will utilize these known properties of Acadian French to develop measurements of the extent to which a corpus exhibits Acadian French properties and could, therefore, be considered to be Acadian French. These properties can be categorized into three parts: Acadian properties, Brayon properties, and Chiac properties. Acadian properties are general tendencies of Acadian French. For example, a token count of known Acadian terms acquired from dictionaries would be an indication [55, 22, 30]. Both Brayon and Chiac are also noted to use English-borrowed verbs [4, 11]. However, we can also measure Brayon and Chiac-specific properties, such as known Brayon words [59] and English code-mixed words for Chiac [11].

Chapter 3

Performance of NLP Tools on Acadian French

In this chapter, we explore our first research question *How well do NLP tools perform on Acadian French text?* by comparing the performance of standard NLP tools when used on standard French and Acadian French. This serves as an evaluation of the need for an Acadian French corpus in the current-day NLP space. If the results indicate that off-the-shelf NLP tools perform more poorly on Acadian French, then an Acadian French corpus would be beneficial as it could help alleviate this by allowing NLP tools to be adapted to Acadian French by training on this corpus.

3.1 Known French Corpora

We first need exemplary Acadian French and non-Acadian French corpora to compare. These corpora need to be known to correspond to these varieties of French. Additionally, these corpora need to be similarly sourced to avoid incomparable documents. We want to do this to avoid comparing two sources held at different standards (e.g. comparing social media posts to French literature). An example of this is the vocabulary typically used in different sources (e.g., abbreviations, informal lan-

guage). Having different vocabulary could affect the results. We chose to solve this by creating corpora from example sentences found within dictionaries of the target French varieties.

The first corpus is made from the Acadian French dictionary [22]. We obtained an electronic version of this dictionary. This meant we could retrieve all the sentences from that dictionary with a basic regex pattern.

The second corpus, the non-Acadian French dictionary, was easier to acquire. Wiktionnaire is a French online dictionary.[1] Using Wiktionnaire, we search for entries and scrape example sentences from them. We use GNU Aspell's list of French words for these search words.[2]

To process these corpora, we apply exact deduplication (Section 4.2.5), followed by near-deduplication (Section 4.2.6), and tokenization (Section 4.2.7). Deduplication refers to removing similar sentences within the corpus. Tokenization refers to segmenting text into a sequence of words. Table 3.1 shows the size of both corpora.

Table 3.1: Number of tokens within each known French corpus.

Corpus Name	Number of Tokens
AcadianDictionary	69,234
Wiktionary	2,627,362

3.2 Masked Language Model Testing

Our first task is mask prediction [48]. We replace a token with a non-informative token (also known as a mask) in a given sentence and calculate the accuracy at which a language model can predict the correct token and the perplexity of the predictions. This process is referred to as *masking a token*. We iterate through each token of each

[1] `https://fr.wiktionary.org/`

[2] `http://aspell.net/`

sentence within the corpus, masking each token in turn, and examine the model's prediction for that token. From that, we count the instances in which the model correctly predicts the target token within its top X most likely predictions. We refer to this as *Accuracy at X*. Additionally, we calculate the perplexity of the task by using Equation 3.1, where N is the total number of words, and $p(w_i)$ is the probability of the word at index i. Perplexity measures the confidence of a language model's predictions. The lower the perplexity, the better the model predicts the text.

$$\text{Perplexity} = -\frac{1}{N}\sum_{\text{i=1}}^{N}\log p(w_i) \tag{3.1}$$

In 2018, Bidirectional Encoder Representations from Transformers (referred to as BERT) became one of the NLP community's most popular models [25]. It was so popular that BERT spun off multiple model variants following its arrival. In 2019, a robustly optimized BERT (referred to as RoBERTa) was created [43]. Soon after, a multilingual version of RoBERTa was created, called XLM-RoBERTa [20]. XLM-RoBERTa is trained on over 2 terabytes of CommonCrawl data, combining 100 languages into one training set. The model was trained to be a multilingual masked language model. Because of this, it is a suitable model to test masked language modelling on both standard French and Acadian French.

Due to size restrictions, we need to shorten the Wiktionary corpus. We randomly sample 4,009 sentences from both corpora. We chose this number because AcadianDictionary has 4,009 total sentences. The tokens within the corpora are 69,234 within AcadianDictionary and 109,720 within Wiktionary. The results are shown in Table 3.2. The columns follow the pattern mentioned above where *Accuracy at 1, 5, or 10* represents the ratio in which the model correctly predicts the target token within its top 1, 5, or 10 tokens predicted.

We also used the Chi-Square test to calculate the p-value and see if the comparisons were statistically significant. To perform a Chi-Square test, we turn the data into

a 2x2 matrix, where the rows are the comparable corpora and the columns are the tokens predicted properly and not predicted properly. The Chi-Square p-values can also be found in Table 3.2. We note a statistical difference given a p-value < 0.05. (No p-value was calculated for the perplexity.)

Table 3.2: XLM-RoBERTa Mask Prediction Results

Corpus	**Accuracy at 1**	**Accuracy at 5**	**Accuracy at 10**	Perplexity
AcadianDictionary	0.382	0.546	0.589	8.129
Wiktionary	0.419	0.586	0.629	7.935
Chi-Square ($p < 0.05$)	6.632e-51	5.755e-59	5.033-62	

As we can see, there seems to be a clear indication that XLM-RoBERTa performs better on standard French text than Acadian French text. All measures of accuracy show a statistically significant difference in performance, always struggling with Acadian French, each of which has roughly a 4 percentage point difference. The perplexity paints a similar picture with a higher score for the Acadian French corpus.

3.3 POS Tagging

Part-Of-Speech (POS) tagging has become a staple in modern NLP pipelines. For example, Stanza requires it to perform lemmatization and create dependency trees [56]. As such, demonstrating the effects of Acadian words on POS tagging could indicate a need for an Acadian corpus to train the models. The Stanford NLP pipeline is a commonly used pipeline that offers multiple tools, including POS tagging. When faced with unknown words, it follows the standard convention and tags the word with X.[3] Table 3.3 shows the number of X tagged tokens within both AcadianDictionary and Wiktionary. It also shows how many sentences in each corpus contain at least one token tagged as X. This is relevant to know because a POS tag error in a sentence could lead to errors in systems that rely on POS tags, such as a dependency

[3] `https://universaldependencies.org/u/pos/X.html`

parser. Table 3.4 shows the top 20 tokens tagged as X within both corpora.

We also used the Chi-Square test to show if the differences are statistically significant. This meant transforming the data into a 2x2 matrix indicating the number of tokens tagged X and the number of tokens not tagged X for both corpora.

Table 3.3: Tokens tagged as Not Found (X) by Stanza

Corpus	**Tok. with *X* tags**	**Rel. Freq.**	**Sentences containing *X* tag**	**Rel. Freq.**
AcadianDictionary	464	0.6701	80	1.9955
Wiktionary	12,790	0.4868	2855	2.9683
Chi-Square	4.6946e-10		0.0051	

Table 3.4: Top 20 Tokens Not Found

Corpus	**Top 20 most frequent tokens tagged *X* in each corpus**
AcadianDictionary	etc., ben, i, in, sus, a, pi, chi, or, to, and, you, al, i', qqn, on, the, as, of, HA
Wiktionary	etc., sic, oh, b, ii, the, a, s., in, l., new, iii, york, i, iv, trump, ben, van, of, paul

We note that both the comparison of tokens tagged not found and sentences containing at least one token tagged not found are statistically different. More tokens are affected, relative to size, in Acadian French. This is expected as the POS tagger has not specifically been trained on Acadian French words. However, more sentences in the standard French corpus relative to size include a token tagged X. This makes sense when considering what tokens were tagged X. Eleven out of the top 20 not found tokens in Wiktionary are nouns, initials (e.g. *S.*, *L.*), and titles (e.g., *Henry III*, *Charles IV*). As such, those tokens would likely not overlap with each other in the same sentence. When calculating the ratio at which tokens tagged X appeared per sentence, Wiktionary averages 4.48 tokens tagged with X per sentence containing at least one token tagged with X, whereas Acadian French has 5.8.

3.4 Summary of the performance of NLP Tools on Acadian French

Both Chapter 3.2 and Chapter 3.3 result in significantly different results when comparing Acadian French text to standard French text. Chapter 3.2 indicates off-the-shelf NLP tools struggle more at mask prediction when given Acadian French text. Similarly, Chapter 3.3 suggests off-the-shelf NLP tools are unable to tag tokens more often in Acadian French text than standard French text. This suggests that Acadian French text is likely not seen in modern-day models and, therefore, unknown to them. Creating an Acadian French text corpus could help alleviate this issue. It would allow models to train, learn, and finetune themselves on Acadian French and possibly better their performance on this text type.

Chapter 4

Corpus Construction

In our work, we built multiple different corpora based on previous web corpus research. We also implemented a post-processing pipeline to extract valuable information from the corpora. Creating a common pipeline for all corpora ensures each corpus is processed in the same way.

4.1 Acadian Corpus Creation

This section will discuss the target corpora created. We create corpora using domain crawling, social media scraping, and the BootCaT approach.

4.1.1 Domain Crawled Corpus

Our first corpus uses French New Brunswick domain websites in the Common Crawl database. We've named this corpus *CC_NB_Domain*. We focus on New Brunswick as the only domain because it holds the largest portion of Acadians per capita.
Common Crawl's data is publicly available through its index server.[1] Each snapshot is stored under a specific index (e.g., CC-MAIN-2023-14, CC-MAIN-2015-14). The API also allows for some useful query parameters. For example, we can specify a

[1] `https://index.commoncrawl.org/`

URL pattern and the output format. We can search for results in a specific language for non-legacy indexes. Additionally, as the data from a specific index returned can sometimes be quite large, we can use their pagination to iterate and retrieve all the data.

We have retrieved all New Brunswick domain websites. We used all indexes starting at "CC-MAIN-2022-21". This allowed us to query 2324 URLs containing French, including websites that are exclusively French and a mix of French and another language.

Common Crawl's data server contains multiple compressed files containing over 100TiB of data.[2] They store the request, response, and metadata.[3]

Part of the data returned from the index server includes information on the targeted HTML and how to access it without downloading the entire file. This includes the endpoint to find and download the data, the size of the data in bytes, and the offset of the stored data in bytes. We can then download byte-specific parts of the endpoint, allowing only the target to be downloaded. All responses stored within Common Crawl's database are WARC formatted; therefore, there is no need to read an unknown file format. However, the encoding of the HTML stored within the WARC file is not always clear. Common Crawl tries to identify the encoding, but it is not always correct. We, therefore, address this in our post-processing pipeline (Section 4.2).

4.1.2 Social Media Corpora

Our second corpus is a social media-based corpus. We use Reddit as it offers accessible APIs to collect data while specifying what type of data to gather using subreddits. We use a specific subreddit, r/acadie, which includes discussions of Acadian-related jokes and discussions. The name of the corpus will be the same as this subreddit.

[2] https://data.commoncrawl.org/

[3] https://commoncrawl.org/the-data/get-started/

Reddit's data can be harvested through a series of API calls to the Reddit servers. However, we chose to use a Python wrapper called PRAW.[4] This wrapper streamlines the requests made to the server to retrieve posts and their comments within Reddit. PRAW allows us to search for a specific subreddit and it gets all the subreddit's posts and each comment within them as well. We combine the post's title, description, comments, and replies into one large text file. We treat each post as one document, similar to one website's HTML from the other corpora. The data is already in UTF-8 text format, so it skips the File Encoding (Chapter 4.2.1) and HTML Text Extraction (Chapter 4.2.2) steps of the post-processing pipeline.

4.1.3 BootCaT Corpora

Our last corpus is our BootCaT corpus. BootCaT works by using online search engines to generate adequate URLs for scraping. It uses tuples of words (often three) within its search. Our first challenge with this approach was that searching for multiple Acadian French words greatly increased the chance to retrieve an online Acadian French dictionary or glossary, as opposed to Acadian French text. To combat this, we combine one Acadian French word, the target word, with two medium-frequency French words from the Aspell dictionary and mandate the target word be within the results.[5] This ensures that the target word would be within the resulting websites while avoiding online dictionaries.

Our second challenge was ensuring a high quantity and quality of data within the resulting URLs. To address this, we only use Acadian French words previously seen in the other Acadian French corpora that we build (i.e., *CC_NB_Domain* and *r/acadie*). The name of this corpus is *BootCaT*.

We used the BootCaT Frontend wizard to create the corpus.[6] We used 45 randomly

[4] `https://praw.readthedocs.io/en/stable/`

[5] `http://aspell.net/`

[6] `https://bootcat.dipintra.it/`

generated tuples. We restricted the search to only HTML documents and filtered adult material. A maximum of 10 URLs were returned per tuple. The wizard then queried and downloaded all URLs retrieved from the Google search engine. The resulting HTML gets placed within our pipeline for post-processing. We chose a small number of tuples to not potentially break Google's terms of service. At the time of writing, we took all necessary steps to ensure that we did not break any terms of service.

4.2 Corpus Processing Pipeline

This section describes the corpus processing process, which takes in HTML and outputs a corpus ready for analysis. This pipeline was inspired by previous work [28, 9, 8].

4.2.1 File Encoding

Our first problem to address is file encoding. HTML files may have different encodings (e.g., UTF-8, Latin-1). Therefore, our task is to re-encode the file using a project-wide encoding standard. We chose UTF-8 as the standard since it extends ASCII. We use the Chardet Python library to guess the original encoding of each HTML file.[7] Chardet receives the raw bytes from the file, guesses what encoding the bytes use, and returns a corresponding confidence rate alongside its response. Anything below a confidence of 0.6 gets removed from the corpus to not pollute the corpus with incorrectly decoded data. The rest gets decoded into a string, based on the encoding Chardet found, and then re-encoded and decoded to UTF-8.

[7] `https://pypi.org/project/chardet/`

4.2.2 HTML Text Extraction

HTML content often has boilerplate content that would create noise within our corpus. HTML boilerplate content refers to repeated code with little change seen throughout different sites; the most common boilerplate content are navigation links, headers, and footers. Therefore, we filter the boilerplate while extracting the text from the HTML. Justext accomplishes both simultaneously [54]. It takes in HTML and outputs clusters of text as paragraphs. The clusters are created based on the HTML tags that surround the content. These paragraphs can then be flagged as boilerplate. Patterns indicate if clusters should be tagged as boilerplate; for example, clusters with little content and links are always boilerplate.[8] We utilize these features to extract the text within the HTML and filter unwanted noise.

4.2.3 Language Filtering

There's no certainty that the text will be French. We, therefore, verify that it is before moving forward. We use langid.py.[9], a Python language classifier. Langid.py classifies documents based on the primarily seen language. It was trained on 5 corpora and can identify 97 languages. Then, based on the trained corpora, it tries to match the input with the languages seen based on patterns identified within the languages. For our purpose, we are looking for primarily French documents and reject all other languages. We look at the entirety of a document within our corpora, and if langid.py's highest confidence is French, then we accept the document as a whole. Even though Acadian French, specifically Chiac (Chapter 2.4), overlaps with English, it contains mostly French words and thus would be flagged as French by langid.py.

[8] `https://corpus.tools/wiki/Justext/Algorithm`

[9] `https://github.com/saffsd/langid.py`

4.2.4 Text Normalization

French text uses accents, which can be represented in numerous ways on the web. Different UTF-8 characters could represent two identical-seeming accents. For example, *è* could be represented by its Latin letter (U+00E8), its Cyrillic letter (U+0450), or by combining the letter *e* and a grave accent (U+0060). This could lead to a scenario where comparing identical seeming strings could lead to a mismatch. Because of this, we need to standardize our corpora.

Unfortunately, comparing Latin to Cyrillic letters is quite difficult. We could strip all accents by temporarily transforming the string to ASCII, transforming it to its closest alphabetical match. However, that would remove all accents and make comparisons between languages difficult. For example, *même* ('same'), a French word, would equate to *meme*, a word in English. Similarly, words could change in meaning. For example, *mangé* ('eating') would equate to *mange* ('eat') which are two different tenses of the same verb. Examples of Cyrillic letters were not found within corpora. Because of this, we chose to ignore this edge case.

The most likely common scenario would be the combination of the letter and its symbol. To address this, we normalize the accents within the French language. The easiest way to do this is to use Unicodedata.[10] Unicode normalization handles two types of normalization: compatibility equivalence and canonical equivalence. Canonical equivalence normalizes the characters. This means it normalizes combined character sequences that form another character. We use this to normalize the corpus text.

Additionally, we apply case folding to our text. This is done to remove duplicated tokens within a corpus vocabulary. This allows for a more accurate token count and helps identify instances of tokens within text.

[10] `https://docs.python.org/3/library/unicodedata.html`

4.2.5 Exact Deduplication

Web paragraphs and documents could have repeated content. Intradocument repetition (paragraph-level) could include titles (e.g. "Introduction," "Description") and repeated dialogue (e.g. repetition of the question within its reply). In contrast, cross-document (document-level) includes terms of services and templated text (e.g. *Mot de la direction:* ('Word from management:'), *À l'Ecole secondaire Népisiguit, nous unissons nos efforts afin:* ('At Népisiguit Secondary School, we are joining forces to:')). Repeated content in a corpus hinders analysis as it emphasizes itself and thus needs to be removed. Hashing can be used to deal with this. Hashing refers to mapping a string of characters to a shorter value, often an integer. We hash all paragraph objects and scan for duplicates in the current document and the entire collection of documents. If we scan an identical hashing, we remove one of the copies. This ensures no document contains identical paragraphs.

4.2.6 Near Deduplication

Similarly, we do not want near-identical data. We are interested in the structure of words and sentences within corpora; thus, near-identical sentences do not give new and meaningful value to a given corpus. An example within our social media corpora is that Reddit users sometimes copy-paste part of the dialogue they reply to. This leads to instances of near identical text within the corpus: *[...] parce qui croit que* ***le NB existe parce que le Quebec lui a accorder*** *ou que le Labrador [...]* ('because who believes that **the NB exists because Quebec granted it** or that Labrador') and >***le NB existe parce que le Quebec lui a accorde*** ('> **the NB exists because Quebec granted it**'). These sentences are not identical; however, the corpus gains very little from adding the second sentence.

To remove near-duplicate sentences, we use PyOnion.[11] PyOnion is a Python version

[11] `https://pypi.org/project/pyonion/`

of the Onion corpus deduplication tool [54]. It uses n-grams to dissect the sentence into multiple parts (of n tokens) and compares the entirety of the n-grams of a given paragraph object (Chapter 4.2.2) to another. Then, it rejects all paragraph objects above a set threshold of overlap. We aimed to be conservative in our choice to include content where possible, so we opted to lean towards strict parameters for detecting near de-duplicates. We chose an n-gram of 5 tokens with a threshold of 0.25, as previous research noted that n-grams of 5, 7, and 10 tokens were common [54]. Our threshold of 0.25 comes from the common options of 0.5, 0.25 and 0.1 [54]; however, 0.1 had an unusable amount of data loss, so we opted for 0.25.

4.2.7 Processing the Result

Lastly, we ready the corpora for analysis. To accomplish this, we use Stanza, a Python library that pipelines the Java Stanford CoreNLP to Python [56].[12] We use Stanza's tokenization for one final filtering of sentences of absurd length (250+ words since sentences of such length are often noise) and English-only sentences since those are most likely quotes or titles of articles. The remaining data re-enters the Stanza pipeline for tokenizing, part-of-speech tagging, lemmatization, and dependency parsing. Table 5.1 showcases the processing results, including token and document count.

While the definition of tokens is identical throughout all corpora, documents in Common Crawl and BootCaT corpora refer to HTML pages. Documents in the context of Reddit corpora refer to posts (and their comments).

[12] `https://stanfordnlp.github.io/CoreNLP/`

Table 4.1: Number of tokens and documents within the target corpora.

Corpus Name	Tokens	Documents
BootCaT	1,208,629	240
CC_NB_Domain	376,668	2324
r/acadie	56,258	801

Chapter 5

Corpus Analysis

This chapter discusses how we will measure how "Acadian" a French text is. We will start by establishing our reference corpus (Chapter 5.1), then by calculating the keyword frequency of our corpora to examine the content within them (Chapter 5.2). Since there is no existing method to determine whether a corpus contains Acadian French, we discuss measures that could help to identify this (Chapter 5.3). We validate these measures by applying them to two corpora of known origin, one Acadian French corpus and another non-Acadian French corpus (Chapter 5.4). Lastly, we will show the results (Chapter 5.5) of comparing the corpora we constructed (Chapter 4.1) to our reference web corpus, frWaC (Chapter 5.1).

5.1 Reference Corpus

This chapter will discuss the reference corpus selected throughout this research. A reference corpus is used as a point of comparison for the targeted corpus. Additionally, our reference corpus must be sourced similarly to our target corpora. This ensures that the source of the content (e.g., web, dictionary) does not affect the results.

We require another web-based corpus for reference. This is because our corpora

stem from the web, so we must compare them to another corpus built from the web. Different media (e.g., web, books, dictionaries) could exhibit different Acadian French properties due to their formality, authors, and standards. Therefore, requiring the reference corpus to be built from the web would ensure that any Acadian French findings within our target corpora would be based on their content and not on the properties of the web itself. As such, we need to find a known French web corpus. This, however, has been previously completed. The WaCky initiative created numerous domain-based corpora, one of which used the .fr domain [10]. Additionally, all their corpora are publicly available for download on their website.[1] We use frWaC as a reference corpus for all our corpora for those reasons. frWaC comes in three sizes: 1 million, 10 million, and 100 million. We chose 10 million since we wanted it to be larger than our biggest target corpus, while not too large to avoid requiring large computational resources to process the corpus. frWaC's data is already text formatted, so the post-processing required is only exact de-duplication (Chapter 4.2.5), near de-duplication (Chapter 4.2.6) and tokenization (Chapter 4.2.7).

Table 5.1: Number of tokens and documents within the frWaC corpus.

Corpus Name	Tokens	Documents
frWaC	11,411,075	16000

5.2 Keyword Analysis

This section showcases the keywords within our built corpora. This will give us an approximation of the type of content each corpus contains and how it differs between corpora.

Comparing the frequency of tokens found in corpora is a means to gather clues about

[1] `https://wacky.sslmit.unibo.it/`

what kinds of text appear most frequently within them. Researchers concluded that utilizing a word's frequency ratio in two corpora can lead to understanding the difference in content between the two corpora [38]. Their equation is shown in Eq. 5.1, where w is the word frequency within a corpus, N is the total number of tokens within the corpus, s is the standard frequency ratio (often one million), and c is a constant that dictates the focus of the equation (smaller values focus on obscure words, while higher values focus on common words). The ratio at which one keyword is more frequent in the target corpus than in the reference corpus can be calculated by Eq. 5.2.

$$\text{KeywordFrequency}(w) = \frac{w * s}{N} + c \tag{5.1}$$

$$\text{KeywordRatio}(w) = \frac{\text{KeywordFrequency}_{\text{focus}}(w)}{\text{KeywordFrequency}_{\text{ref}}(w)} \tag{5.2}$$

We calculate the keyword ratio of each newly built web corpus with frWaC as their reference corpus, using a constant of 100, the default value suggested by researchers [38]. Table 5.2 showcases the top 20 keywords within each target corpus.

Table 5.2: Top 20 keywords in each corpus with frWaC as the reference corpus.

Target Corpus	Keywords
BootCaT	u, r, e, o, t, i, québec, n, canada, autochtones, d, québécois, yolanda, caribou, p, steven, -là, l'on, canadiens, q
CC_NB_Domain	nouveau-brunswick, officielles, élèves, canada, moncton, langues, école, élève, province, scolaire, commissaire, -vous, linguistiques, francophone, coucher, fredericton, langue, salle, soins, assurance
r/acadie	acadien, acadie, acadiens, **, acadienne, chiac, pis, *, francophones, québec, the, mot, nouveau-brunswick, francophonie, and, canada, to, acadiennes, québécois, langue

Before the analysis, something to note is that the BootCaT corpus has single character keywords. This indicates that there is likely a mismatch between the tokenization of this corpus and the reference corpus, frWaC. However, there are still keywords standing out to discuss.

BootCaT has, first and foremost, single-character tokens appearing a total of ten times. These tokens, in order, are: *u*, *r*, *e*, *o*, *t*, *i*, *n*, *d*, *p*, and *q*. They seem to originate from recorded discussions posted on the web, where these tokens are the initials of the speakers. Additionally, the keywords include *québec*, *canada*, *autochtones* ('indigenous'), *québécois*, and *canadiens*. This suggests a large amount of Canadian data found, focusing on Québec. Most keywords in the corpus suggest this hypothesis, even *caribou* ('reindeers'), as it was found in a Canadian article about them. *yolanda* and *steven* come from a transcript of a conversation between two medical practitioners. The remaining words, *-là* and *l'on*, are seen in standard French.

CC_NB_Domain's keywords are oriented toward the education sector. This is because the education sector of the province of New Brunswick uses the domain *nbed.nb.ca*, which belongs within the *.nb.ca* domain. The data reflects this with keywords such as *élèves* ('students'), *école* ('school'), *élève* ('student'), *scolaire* ('scholar'), *salle* ('classroom'). Additionally, in the past, New Brunswick municipal sites were hosted within the *.nb.ca* domain. An example of this within the corpus is *village.stantoine.nb.ca*. This reflects location and residential-based keywords such as *moncton*, *coucher* ('sleep'), and *fredericton*. Lastly, documents from the domain *languesofficielles.nb.ca* were prevalent within the corpus, which explains keywords such as *officielles* ('official'), *langues* ('languages'), *commissaire* ('commissioner'), *linguistiques* ('linguistics'), *francophone*, and *langue* ('language'). Location-based words like *nouveau-brunswick*, *canada*, and *province* make sense based on the domain. *soins* ('care') and *assurance* ('insurance') are seen throughout many documents; however, they are mostly seen in a health insurance document. And lastly, *-vous* is a word seen in standard French.

r/acadie seems to include linguistic-based keywords. r-acadie focuses on Acadian French-related data, with *acadien*, *acadie*, *acadiens*, *acadienne*, *chiac*, and *acadiennes* all being directly related to Acadian French. Acadian French dialect words appear

as well with *pis*, *the*, *and*, and *to*. *francophones*, *francophonie*, *langue* directly relate to the French language. *québec*, *nouveau-brunswick*, *canada*, and *québécois* are all Canadian-oriented tokens. *mot* ('word') was seen in posts discussing words and the context in which you could find them. Lastly, ** and * are stylistic indicators found within Reddit's text (****bold text**** and **italic text*).[2]

These findings suggest that all of our corpora are Canadian-French oriented in their content. However, some lean towards an untargeted Quebec French rather than the targeted Acadian French.

5.3 Structure of Acadian French

To analyze if a corpus contains Acadian French, we first need to know the linguistic characteristics of Acadian French text. Since there are multiple types of Acadians within New Brunswick [11], we need to consider all of them within our analysis. As such, we divided this chapter into three parts to better represent the different Acadians. These parts are *General Acadian French*, *Brayon French*, and *Chiac French*. The *General Acadian French* section measures characteristics that could be seen throughout New Brunswick, overlapping both Brayon French and Chiac French. The *Brayon French* section measures characteristics that can be mostly seen in the northwest regions of New Brunswick. Lastly, the *Chiac French* section measures characteristics mostly found in the southeast regions of New Brunswick. Throughout this chapter, let N be the total number of tokens (e.g., sets of characters that provide a useful semantic unit) and V be the total number of types (e.g., the total number of different tokens) within a corpus. To compute these measures, a tokenizer, part-of-speech tagger, lemmatizer, and dependency parser are required (Chapter 4.2.7). A summary is presented in Table 5.3, which states all 22 measures of Acadian French and how they are computed.

[2] `https://support.reddithelp.com/hc/en-us/articles/360043033952-Formatting-Guide`

5.3.1 General Acadian French Structure

In this section, we will discuss various characteristics of the structure of most Acadian French texts. This means that this section would apply to most Acadian French within the region of New Brunswick.

5.3.1.1 Acadian French Tokens

Acadian French has a unique vocabulary composed of words seen only within the dialect [22]. These words would not appear in a standard French dictionary. We chose the Acadian Dictionary [22] to source all Acadian French words. The Acadian Dictionary is a dictionary dedicated to listing every Acadian French term and does not hold any definitions of words outside the scope of Acadian French. Therefore, all words in that dictionary are found in Acadian French. By counting token instances of words within that dictionary, we calculate the ratio at which Acadian French tokens appear instead of non-Acadian French tokens. We calculate this with the following equation:

$$\text{Acadian French Tokens Ratio} = \frac{\text{Number of Acadian French tokens}}{\text{N}} \tag{5.3}$$

5.3.1.2 Acadian French Types

We also count the number of Acadian French types within the corpus by again leveraging the Acadian dictionary [22]. This indicates the diversity of Acadian French vocabulary found in a corpus. We can calculate this using the following equation:

$$\text{Acadian French Type Ratio} = \frac{\text{Number of Acadian French types}}{\text{V}} \tag{5.4}$$

While similar, Equation 5.3 and Equation 5.4 measure completely different things. In Equation 5.3, we calculate the total number of Acadian French tokens found in a

corpus. This is a great indicator since Acadian French text would see more Acadian French tokens than non-Acadian French text. However, if a corpus repeats only one Acadian French token many times throughout, it would not be considered Acadian French. Therefore, only measuring the number of times Acadian French tokens appear is insufficient. Equation 5.4, on the other hand, measures the diversity of Acadian French vocabulary in a corpus. This is also a great indicator of Acadian French. However, there exist websites that describe Acadian French words in standard French. Examples of this would be online Acadian French glossaries and word banks.[3][4] These types of websites would inflate the number of Acadian French types seen. Therefore, the text extracted, while having a high Acadian French Type number, would mainly contain standard French. For these reasons, having both measures, Equation 5.3 and 5.4, is extremely valuable as they complement one another.

5.3.1.3 Auxiliary Avoir

In French, verbs can be accompanied by another verb to express a certain mood, voice, or tense. These accompanied verbs are known as *auxiliary verbs*, and only two verbs can be auxiliary verbs: *avoir* and *être*. Acadians often use the auxiliary *avoir* when the auxiliary *être* would be appropriate in standard French [11]. An example would be *je m'ai baigné* ('I went swimming') as opposed to the standard French *je me suis baigné*. As such, one way to measure this is by calculating the ratio at which the auxiliary *avoir* appears over all auxiliaries. We can calculate this as follows:

$$\text{Auxiliary Avoir Ratio} = \frac{\text{Number of auxiliary } \textit{avoir} \text{ tokens}}{\text{Total number of auxiliary tokens}} \tag{5.5}$$

[3] https://sagouine.com/fr/component/content/category/17-glossaire

[4] http://www.acadian-home.org/acadian-words.html

5.3.1.4 Acadian Conjunctions

Some non-standard conjunctions appear in Acadian French [68]. These conjunctions include *pis*, *ben*, and *so*. The entire set of these conjunctions is listed in previous work [68]. As such, measuring the ratio at which these conjunctions appear over all other conjunctions (e.g., tokens tagged *CCONJ* by our POS tagger) could indicate if a text is Acadian.[5] We can calculate this as follows:

$$\text{Acadian Conjunctions Ratio} = \frac{\text{Number of Acadian conjunction tokens}}{\text{Total number of conjunction tokens}} \tag{5.6}$$

5.3.1.5 Acadian Prepositions

It was noted that a few prepositions accompany infinitive verbs more often in Acadian French than others (e.g., *tout ce*, *que c'est*, *which que*) [68]. That research showcased a table of all prepositions seen through different varieties of French [68]. We count those listed under Acadian French and compare them to the total number of all listed prepositions. We can calculate this using the following equation:

$$\text{Acadian Preposition Ratio} = \frac{\text{Number of Acadian preposition tokens}}{\text{Total number of preposition tokens}} \tag{5.7}$$

5.3.1.6 Questions Containing *ti*

Acadian French often adds *ti* as a marker of questions [39, 52, 63, 11]. An example of this would be "Il va ti partir bientôt?" ('Will he leave soon?') [63]. As such, measuring the ratio at which *ti* appears within questions could indicate if a text is Acadian. We assume that any sentence ending in a question mark is a question and calculate this using the following equation:

[5] `https://universaldependencies.org/u/pos/CCONJ.html`

$$\text{Questions Containing } ti \text{ Ratio} = \frac{\text{Num. of question contaning the token } ti}{\text{Total number of questions}} \quad (5.8)$$

5.3.1.7 English-Borrowed Acadian French Verb Tokens

Acadian French includes *-ing* suffixed English verbs, such as *parking*, *eating*, *biking* transformed to French *-er* suffixed verbs, *parker*, *eater*, *biker* [50, 12, 11, 4]. As such, measuring the ratio at which these forms occur over the instances of all verbs could indicate if a text is Acadian. To do this, we take an English verb forms dictionary and filter for *-ing* suffixed verbs.[6] We then replace the suffix *-ing* with *-er* and check that it is not in the French dictionary. We do this to ensure that there is no overlap between standard French and this measure so as not to potentially inflate the results. This creates an English-borrowed Acadian French verb list. We can calculate this measure using the following equation:

$$\text{English Verbs Ratio} = \frac{\text{Number of English-borrowed verb tokens}}{\text{Total number of verb tokens}} \quad (5.9)$$

5.3.1.8 English-Borrowed Acadian French Verb Types

Similarly, we count the number of types of English-Borrowed Acadian French verbs within the corpus [11, 4]. This shows the diversity of English-borrowed verbs used within the corpus. As such, measuring the ratio at which unique derivations occur over the instances of all verb types could indicate if a text is Acadian French. We use the same English-borrowed Acadian French verb list as for Equation 5.9. We calculate this ratio using the following equation:

[6] `https://github.com/monolithpl/verb.forms.dictionary`

$$\text{English Verbs Type Ratio} = \frac{\text{Number of English-Borrowed verb types}}{\text{Total number of verbs types}} \tag{5.10}$$

5.3.1.9 Instances of *point*

Acadian French uses *point* as a form of negation [11, 53]. As such, measuring the ratio at which *point* appears over *pas* as a negation could indicate if a text is Acadian French. To do this, we consider all instances in which either *point* or *pas* are found in the text tagged *ADV* (adverb) by the part-of-speech tagger and *advmod* (adverbial modifier) by the dependency parser.[7][8] We then calculate this using the following equation:

$$\textit{point} \text{ Ratio} = \frac{\text{Number of } \textit{point} \text{ adverb tokens}}{\text{Number of } \textit{point} + \textit{pas} \text{ adverb tokens}} \tag{5.11}$$

5.3.2 Brayon French Structure

In this chapter, we will be discussing various Brayon-specific characteristics that could be found within the Brayon French text.

5.3.2.1 Brayon French Tokens

Brayon French uses words unique to this variant of Acadian French [59, 60]. Examples of these words are: *Bagosse* ('East Edmunston'), *balancine* ('swing'), *dewères* ('homework'). These words would not appear in a standard French dictionary, and most would not appear in the general Acadian French dictionary. We chose the Brayonnaire to source all Brayon French words [59, 60]. By counting token instances of words from that dictionary in a corpus, we calculate the ratio at which Brayon

[7] `https://universaldependencies.org/u/pos/ADV.html`

[8] `https://universaldependencies.org/u/dep/advmod.html`

French tokens appear instead of non-Brayon French tokens. We can calculate this ratio using the following equation:

$$\text{Brayon French Tokens Ratio} = \frac{\text{Number of Brayon French tokens}}{\text{N}} \tag{5.12}$$

5.3.2.2 Brayon French Types

Similarly, we also count the number of Brayon French types within the corpus using the Brayonnaire dictionary [59, 60]. This indicates the diversity of Brayon French types found in a corpus. We can calculate this using the following equation:

$$\text{Brayon French Type Ratio} = \frac{\text{Number of Brayon French types}}{\text{V}} \tag{5.13}$$

5.3.2.3 Brayon French Expressions

Brayon French also contains unique expressions found within the Brayonnaire dictionary [59, 60]. Examples of these are: *Chu brûlé bin tight* ('I'm totally tired out') and *Zip ton coat, t'as toute la falle à l'air* ('Zip up your coat, your throat's out in the cold'). The presence of such expressions could indicate the usage of Brayon French within a corpus. To calculate this, we count the sentences where one or more Brayon expressions occur and compare the results to the total number of sentences within the corpus. We calculate this ratio with the equation that follows:

$$\text{Brayon Expressions Ratio} = \frac{\text{Num. of sentences w/ a Brayon expression}}{\text{Total number of sentences}} \tag{5.14}$$

5.3.2.4 Adverbs ending in *-eux*

One identifying characteristic of Brayon French text is the usage of adverbs with the suffix *-eux* (e.g., *lamenteux* ('lamentable'), *brailleux* ('crier')) [4]. Brayon French

tends to use these adverbs more frequently than non-Brayon French. As such, the ratio at which adverbs with the suffix "-eux" appear over any other adverb could indicate Brayon French text. We can calculate this as follows:

$$\text{Brayon Adverbs Ratio} = \frac{\text{Number of adverb tokens with the suffix } \textit{-eux}}{\text{Total number of adverb tokens}} \tag{5.15}$$

5.3.3 Chiac French Structure

In this section, we will be discussing Chiac-specific characteristics that could be found within Chiac French text.

5.3.3.1 English Tokens

Chiac French often uses English words [12, 11]. Token instances of English words within a corpus can indicate that the corpus includes Chiac French text. For a list of English words, we use GNU Aspell's English dictionary.[9] To eliminate the instances of words that can be found in both French and English, we count the words that only appear in the English dictionary and not any French dictionary (neither standard French nor Acadian French). For our standard French dictionary, we use GNU Aspell's French dictionary. We turn this into a ratio by comparing the English token count to the entire token count of a corpus. We can calculate this as follows:

$$\text{English Tokens Ratio} = \frac{\text{Number of English tokens}}{\text{N}} \tag{5.16}$$

[9] http://aspell.net/

5.3.3.2 English Types

Similarly, we also count the number of English types within the corpus [12, 11]. This indicates the diversity of English words found in a corpus. We can calculate this ratio using the following equation:

$$\text{English Type Ratio} = \frac{\text{Number of English types}}{\text{V}} \tag{5.17}$$

5.3.3.3 3rd Person *-ont* Verbs

Chiac French changes the suffix of 3rd person plural French verbs from *-ent*/*-ant* to *-ont* [11]. An example of this variation would be "Ils mang*ont* bien!" ('They eat well!'). The ratio at which *-ont* suffixed 3rd person plural verbs appear as opposed to their *-ent*/*-ant* suffixed counterparts could be an indication of Chiac text. We can calculate this using the following equation:

$$\text{3rd Person Plural Verb Ratio} = \frac{\text{Num. 3rd person plural } \textit{-ont} \text{ verb tokens}}{\text{Total num. 3rd person plural verb tokens}} \tag{5.18}$$

5.3.3.4 *-ly* Adverb Tokens

Chiac French replaces *-ment* suffixed French adverbs with *-ly* suffixed English adverbs [11]. Examples of this would be replacing *lentment* for *slowly* and *violentment* for *violently*. As such, the ratio at which these English adverbs occur as opposed to their French counterparts could indicate Chiac French text. English adverbs can easily be obtained by counting the instances of English words ending in *-ly*, and French adverbs are obtained by counting the instances of tokens tagged as *ADV* (adverb), that do not appear in the English dictionary, and that end in *-ment*.[10] We can calculate this as follows:

[10] https://universaldependencies.org/u/pos/ADV.html

$$\textit{-ly}\text{ Adv. Token Ratio} = \frac{\text{Num. of }\textit{-ly}\text{ suffixed English adv. tokens}}{\text{Total num. of }\textit{-ly}\text{ Eng. + }\textit{-ment}\text{ Fr. adv. tokens}} \quad (5.19)$$

5.3.3.5 *-ly* Adverb Types

Similarly, we also count the number of *-ly* suffixed English adverb types within the corpus in comparison to the number of *-ment* suffixed French adverb types [11]. This indicates the diversity of Chiac French adverbs found in a corpus. We can calculate this using the following equation:

$$\textit{-ly}\text{ Adv. Type Ratio} = \frac{\text{Num. of }\textit{-ly}\text{ suffixed English adv. types}}{\text{Total num. of }\textit{-ly}\text{ Eng. + }\textit{-ment}\text{ Fr. adv. types}} \quad (5.20)$$

5.3.3.6 Instances of *you know*

Chiac French uses English's *you know* over French's counterparts: *tu sais* and *t'sais* [11]. As such, the number of instances of *you know* compared to French alternatives could indicate Chiac French text. We can calculate this as follows:

$$\textit{You Know}\text{ Ratio} = \frac{\text{Num. of instances of }\textit{you know}}{\text{Total num. of }\textit{you know}\text{ + French alternatives.}} \quad (5.21)$$

5.3.3.7 Instances of *right* adverb

Chiac French uses English's *right* as a replacement for other French adverbs [50, 11]. Counting the instances of *right* tokens compared to other French adverbs could indicate Chiac text. We can calculate this using the following equation:

$$\textit{Right}\text{ Ratio} = \frac{\text{Number of adverbial }\textit{right}\text{ tokens}}{\text{Total number of abverb tokens}} \quad (5.22)$$

5.3.3.8 Instances of *back* adverb

Chiac French uses English's *back* over other French adverbs [11]. Counting the instances of *back* adverb tokens compared to French adverbs could indicate Chiac text. We can calculate this using the following equation:

$$\textit{Back}\ \text{Ratio} = \frac{\text{Number of adverbial } \textit{back} \text{ tokens}}{\text{Total number of abverb tokens}} \tag{5.23}$$

5.3.3.9 Instances of *own*

Chiac French often uses English's *own* instead of French alternatives: *à moi-même* and *propre* [11, 52]. Counting the instances of *own* tokens over standard French alternatives could indicate Chiac text. We can calculate this as follows:

$$\textit{own}\ \text{Ratio} = \frac{\text{Number of } \textit{own} \text{ tokens}}{\text{Total num. of } \textit{own} + \text{French alternatives.}} \tag{5.24}$$

5.3.4 Summary of the Acadian French Measures

In this section, we will summarize all 22 measures of Acadian French. Table 5.3 summarizes all 22 measures into a table format with their corresponding formulas to calculate them. Also, we separated the Table into three parts: Acadian French, Brayon French, and Chiac French. This is done to identify the type of Acadian French found within each corpus. This table format will be consistent throughout the remainder of this thesis.

5.4 Benchmark Comparison

Before we compare our newly built web corpora to our known French web corpora, we need to understand if the measures correctly identify Acadian French text. In this chapter, we compare a known Acadian French corpus, AcadianDictionary (Chapter

Table 5.3: Summary of all Acadian French measures

	Acadian French Measures	**Corresponding Formulas**
Acadian	Acadian French Tokens	$\frac{\text{Number of Acadian French tokens}}{\text{N}}$
	Acadian French Types	$\frac{\text{Number of Acadian French types}}{\text{V}}$
	Auxiliary Avoir	$\frac{\text{Number of auxiliary } \textit{avoir} \text{ tokens}}{\text{Total number of auxiliary tokens}}$
	Acadian Conjunctions	$\frac{\text{Number of Acadian conjunction tokens}}{\text{Total number of conjunction tokens}}$
	Acadian Prepositions	$\frac{\text{Number of Acadian preposition tokens}}{\text{Total number of preposition tokens}}$
	Questions Containing *ti*	$\frac{\text{Num. of question contaning the token } \textit{ti}}{\text{Total number of questions}}$
	English-Borrowed Acadian French Verb Tokens	$\frac{\text{Number of English-borrowed verb tokens}}{\text{Total number of verb tokens}}$
	English-Borrowed Acadian French Verb Types	$\frac{\text{Number of English-Borrowed verb types}}{\text{Total number of verbes types}}$
	Instances of *point* Negation	$\frac{\text{Number of } \textit{point} \text{ adverb tokens}}{\text{Number of } \textit{point} + \textit{pas} \text{ adverb tokens}}$
Brayon	Brayon French Tokens	$\frac{\text{Number of Brayon French tokens}}{\text{N}}$
	Brayon French Types	$\frac{\text{Number of Brayon French types}}{\text{V}}$
	Brayon French Expressions	$\frac{\text{Num. of sentences w/ a Brayon expression}}{\text{Total number of sentences}}$
	Adverbs ending in *-eux*	$\frac{\text{Number of adverb tokens with the suffix } \textit{-eux}}{\text{Total number of adverb tokens}}$
Chiac	English Tokens	$\frac{\text{Number of English tokens}}{\text{N}}$
	English Types	$\frac{\text{Number of English types}}{\text{V}}$
	3rd Person *-ont* Verbs	$\frac{\text{Num. 3rd person plural } \textit{-ont} \text{ verb tokens}}{\text{Total num. 3rd person plural verb tokens}}$
	-ly Adverb Tokens	$\frac{\text{Num. of } \textit{-ly} \text{ suffixed English adverb tokens}}{\text{Total num. of } \textit{-ly} \text{ Eng. + } \textit{-ment} \text{ Fr. adverb tokens}}$
	-ly Adverb Types	$\frac{\text{Number of } \textit{-ly} \text{ suffixed English adverb types}}{\text{Total num. of } \textit{-ly} \text{ Eng. + } \textit{-ment} \text{ Fr. adverb types}}$
	Instances of *you know*	$\frac{\text{Num. of instances of } \textit{you know}}{\text{Total num. of } \textit{you know} + \text{ French alternatives.}}$
	Instances of *right* adverb	$\frac{\text{Number of adverbial } \textit{right} \text{ tokens}}{\text{Total number of abverb tokens}}$
	Instances of *back* adverb	$\frac{\text{Number of adverbial } \textit{back} \text{ tokens}}{\text{Total number of abverb tokens}}$
	Instances of *own*	$\frac{\text{Num. of } \textit{own} \text{ tokens}}{\text{Total num. of } \textit{own} + \text{ French alternatives.}}$

3.1), to a known standard French corpus, Wiktionary (Chapter 3.1), to determine if the Acadian French measures can correctly identify the Acadian French corpus. To correctly identify Acadian French text, the measures must collectively indicate that the AcadianDictionary corpus is Acadian French. Suppose the measures can identify the AcadianDictionary corpus as such. If that is the case, then they could identify Acadian French within an unknown French corpus (i.e., a corpus for which we do not know its variety of French such as the three web corpora in Section 4.1). Additionally, we compare AcadianDictionary to frWaC since frWaC will be used as the reference corpus for all three of our newly built web corpora (Chapter 4.1). Knowing how a known Acadian French source compares against the reference corpus is crucial. We will use these results as a point of comparison in Chapter 5.5.

To compare two corpora, we use each measure as a means to identify Acadian French characteristics. An example of a measure, that we will refer to throughout this paragraph, is *Acadian French Tokens* (Chapter 5.3.1.1). To use a measure, we first need to count the number of instances at which the Acadian French characteristic occurred within both corpora. In our example, we count the total number of Acadian French tokens found within both corpora. Afterwards, we count the number of instances where the characteristic could have occurred but did not in both corpora. Following our example, the number of instances where there could have been an Acadian French token is the total number of tokens. However, we want the instances where it could have occurred but did not. Therefore, we would take the total number of tokens and subtract the number of Acadian French tokens. This leaves us with two counts for both corpora and from that, we create a 2x2 matrix, where the rows are the corpora and the columns are the occurrences of the characteristic and lack thereof. For our example, the matrix would be Table 5.4.

Using the 2x2 matrix allows us to calculate if the difference for each measure is statistically significant ($p < 0.05$). Unlike Chapter 3, we use Fisher's Exact Test

Table 5.4: Example 2x2 matrix of the Acadian French Token measure, used to calculate Fisher's Exact Test

	Acadian Tokens	**non-Acadian Tokens**
AcadianDictionary	1,808	67,426
Wiktionary	21,270	2,606,092

[66] instead of Chi-Square Test [62]. Past research suggested using Fisher's for cases where $N < 5$ [19]; however, more recently, it was noted that Chi-square is sufficient for all cases where $N > 0$ [17]. Since it is improbable that all Acadian measures are found within all corpora, Fisher's Exact Test is needed where $N = 0$ would occur. Fisher's Exact Test, while exact for small samples, is known for being conservative at a larger scale [65]. While this does not deter us from using it, it is worth noting as the results are conservative.

Table 5.5 shows the results of these comparisons. The statistically significant results are colour-coded by green and red ($p < 0.05$). Green signifies that the measure was higher in the target corpus (i.e., AcadianDictionary), and red signifies that the measure was equal or higher in the reference corpus (i.e., Wiktionary and frWaC). Yellow results show that the comparison was not statistically significant.

As we can see, AcadianDictionary shows more Acadian French measures than Wiktionary. Most measures correctly identified the Acadian French source, with the exception of two: *Acadian Prepositions* (Equation 5.7) and *English Types* (Equation 5.17). This means, as a collective, these measures can identify an Acadian French text over a standard French text.

On the other hand, when we compare AcadianDictionary to frWaC, it shows fewer significant Acadian French measures than the comparison to Wiktionary. Additionally, when we compare AcadianDictionary to frWaC, it shows more measures were significantly less frequent in the target corpus (i.e., red in Table 5.5) than the comparison to Wiktionary. This means, our measures are less indicative that Aca-

Table 5.5: Comparisons of the baseline target corpus AcadianDictionary to two reference corpora. Green cells indicate the Acadian property is significantly more frequent in the target corpus. Red cells indicate the Acadian property is significantly less frequent in the target corpus. Yellow cells indicate the results were not statistically significant.

	Acadian French Measures	**Wiktionary**	**FrWaC**
Acadian	Acadian French Tokens		
	Acadian French Types		
	Auxiliary Avoir		
	Acadian Conjunctions		
	Acadian Prepositions		
	Questions Containing *ti*		
	English-Borrowed Acadian French Verb Tokens		
	English-Borrowed Acadian French Verb Types		
	Instances of *point* Negation		
Brayon	Brayon French Word Tokens		
	Brayon French Word Types		
	Brayon French Expressions		
	Adverbs ending in *-eux*		
Chiac	English Tokens		
	English Types		
	3rd Person *-ont* Verbs		
	-ly Adverb Tokens		
	-ly Adverb Types		
	Instances of *you know*		
	Instances of *right* adverb		
	Instances of *back* adverb		
	Instances of *own*		

dianDictionary contains Acadian French when compared to frWaC as opposed to Wiktionary. This is important to note for our analysis benchmark.

While we do not have a concrete reason for this, the creation method of frWaC may be a factor. Both AcadianDictionary and Wiktionary were created from samples of example sentences of entries within dictionaries. frWaC, on the other hand, was created from numerous *.fr* domain websites. This means that frWaC could hold more diverse usages of French (since each website could be written in a different variety of French). In fact, frWaC could even include some Acadian French, even though that was not the intended purpose. The uncertainty of the sources of data within frWaC definitely could have impacted these results.

All in all, the 22 measures we built, to identify Acadian French properties, were shown to be capable of identifying Acadian French text. This means we can use these measures to test if our three newly built target web corpora (BootCaT, CC_NB_Domain, and r/acadie) hold Acadian French text.

5.5 Results

In this section, we aim to answer both remaining research questions: *Can we create an Acadian French corpus using previously proposed web-as-corpus methodologies?* and *Is there one of these newly built corpora that surpasses others in quality and quantity?* We will identify if our newly built web corpora contain Acadian French text using the Acadian characteristics described in Chapter 5.3.

We have three target corpora: BootCaT, CC_NB_Domain, and r/acadie. Additionally, we have a fourth corpus acting as a benchmark point of comparison corpus, AcadianDictionary. All four corpora will be compared to the same reference corpus: frWaC. We chose frWaC, a large, commonly used French web corpus, because while it may have traces of Acadian French, representing Acadian French was not

the project's intended goal. We compare four corpora to the same reference corpus, unlike Table 5.5, which compares one target corpus to two reference corpora. Thus, the column will be the varying corpora, i.e. the target corpora.

Table 5.6 shows the results of these comparisons. We compare them in the same way as mentioned in the benchmark (Chapter 5.4), which is by comparing the ratio of each measure in both corpora and using Fisher's Exact test to indicate if the comparison is statistically significant. The statistically significant results are colour-coded by green and red ($p < 0.05$). Green signifies that the measure was higher in the target corpus (i.e., AcadianDictionary, BootCaT, CC_NB_Domain, and r/acadie), and red signifies that the measure was equal or higher to the reference corpus (i.e., frWaC). Yellow indicates that the comparison was not statistically significant.

Table 5.7 shows a summarized version of Table 5.6. It summarizes the total count of Acadian French measures in each category: the number of measures that were higher in the target corpus (green), the number of measures that were equal or lower in the target corpus (red), and the number of measures that were not statistically significant (yellow).

As we can see, all corpora hold more Acadian French tokens and types by amount than our reference corpus, frWaC. This demonstrates that not only do all three newly built corpora show traces of Acadian French, but they also have significantly more Acadian French tokens and types than a French web corpus. On the other hand, websites could be discussing Acadian French words and not properly using them within a sentence. Regardless, each corpus exhibits additional unique attributes aside from that, indicating that they hold Acadian French text.

BootCaT shows identical results in Brayon French as the AcadianDictionary. General Acadian French measures are also very similar to AcadianDictionary except *English-Borrowed Acadian French Verb Tokens* (Chapter 5.3.1.7). Most of the measures not found are for Chiac, yet it still exhibits traces of a few characteristics of

Table 5.6: Comparisons of all target corpora to the frWaC reference corpus. Green cells indicate the Acadian property is significantly more frequent in the target corpus. Red cells indicate the Acadian property is significantly less frequent in the target corpus. Yellow cells indicate the results were not statistically significant.

	Acadian French Measures	**AcadianDictionary**	**BootCaT**	**CC_NB_Domain**	**r/acadie**
Acadian	Acadian French Tokens				
	Acadian French Types				
	Auxiliary Avoir				
	Acadian Conjunctions				
	Acadian Prepositions				
	Questions Containing *ti*				
	English-Borrowed Acadian French Verb Tokens				
	English-Borrowed Acadian French Verb Types				
	Instances of *point* Negation				
Brayon	Brayon French Tokens				
	Brayon French Types				
	Brayon French Expressions				
	Adverbs ending in *-eux*				
Chiac	English Tokens				
	English Types				
	3rd Person *-ont* Verbs				
	-ly Adverb Tokens				
	-ly Adverb Types				
	Instances of *you know*				
	Instances of *right* adverb				
	Instances of *back* adverb				
	Instances of *own*				

Table 5.7: Summary of All Acadian French Measures Compared to frWaC. Green cells indicate the Acadian property is significantly more frequent in the target corpus. Red cells indicate the Acadian property is significantly less frequent in the target corpus. Yellow cells indicate the results were not statistically significant.

Corpus			
AcadianDictionary	8	4	10
BootCaT	9	6	7
CC_NB_Domain	4	5	13
r/acadie	13	1	8

Chiac, such as *English Tokens* (Equation 5.16) and *3rd Person -ont Verbs* (Equation 5.18). BootCaT tied with the benchmark corpus for the highest number of statistically significant Brayon French characteristics found within a corpus. This could be due to the overlap between Brayon and Quebec French. Quebec, a larger province in Canada, potentially has more internet presence and thus would be seen more throughout search engine results. This is also further observed within the keyword analysis (Chapter 5.2) since *québec* and *québécois* are frequent in the corpus. We hypothesize that during corpus construction, a combination of restricting the search by domain and content verification could likely solve this overlap issue.

CC_NB_Domain yielded uncertain results. Fewer than ten measures are significant, and they roughly yield a fifty-fifty split between having a higher frequency in BootCaT (green) or not (red). Measures being insignificant is likely due to size and content. It was restricted to the content of the crawled *.nb.ca* domain. That domain holds the provincial education websites, most of which are written in standard French. This potentially restricted the presence of Acadian French in the corpus. Additionally, unlike some other domains (e.g., *.ca*, *.uk*), *.nb.ca* is quite small. Further work could explore the relative size needed for accurate dialect representation when crawling a domain. We hypothesize that population and internet presence will likely play a major role.

Finally, r/acadie showed the highest number of statistically significant Chiac French properties within a corpus and the highest overall number of statistically significant measures. Additionally, very few measures had a frequency tied to or smaller than in the reference corpus. In fact, r/acadie has no measures in either Brayon or Chiac, in which its frequency is tied to or smaller than in frWaC. Because of this, r/acadie holds more statistically significant measures than our benchmark corpus, Acadian-Dictionary. This means that it likely consists of Acadian French text. Other social media platforms would be worth investigating to see if they yield similar or better

results.

In this chapter, to answer our second research question, we evaluated all three web corpora by calculating their keyword frequency. This showed traces of Canadian French-oriented content were found within all three web corpora. Some even showed traces of Acadian French. Then, we compared the ratio of Acadian French corpus-based statistical measures of all three web corpora to a reference corpus, frWaC. This test also showed that all three corpora have traces of Acadian French text. Therefore, we can conclude that all three corpora hold some Acadian French text. Additionally, to answer our third research question, we compared the corpus-based statistical measures to a benchmark known Acadian French corpus. This showed that BootCaT contains roughly as many characteristics as our benchmark and that r/acadie surpasses our benchmark in the number of Acadian French characteristics present within the corpus. Future work could investigate if adapting the models mentioned in Chapter 3 to Acadian French using these corpora could fix the issues they were having processing Acadian French.

Chapter 6

Conclusion

Acadian French is a minority dialect with no large representative corpus. Creating corpora for dialects is important for improving the performance of NLP models. In this thesis, we evaluated the performance of off-the-shelf NLP tools on Acadian French text. We also created three Acadian French web corpora using the following web-as-corpus creation methods: BootCaT, domain crawling, and social media scraping. Finally, we proposed 22 statistical corpus-based measures stemming from previously researched Acadian French characteristics to compare these newly built corpora to known Acadian French text.

We started by examining our first research question: *How well do NLP tools perform on Acadian French text?*. This served to confirm the importance of our other research questions. Our findings showed that off-the-shelf NLP tools perform worse on Acadian French text than on standard French text. Masked language modelling, results seen in Table 3.2, performed significantly worse on Acadian French. Both accuracy and perplexity supported this conclusion. Moreover, an off-the-shelf part-of-speech model tagged more tokens as X (unassignable) for Acadian French than for standard French, suggesting that current-day NLP tools perform worse on Acadian French text.

Our second research question was *Can we create an Acadian French corpus using previously proposed Web-as-Corpus methodologies?* We created three corpora stemming from known web-as-corpus methodologies. The first was a domain-crawled corpus. We used CommonCrawl's extensive database and scraped New Brunswick websites, making a corpus from the results. The second was a social media-based corpus, where we gathered all posts and comments in the *r/acadie* subreddit. The third and last was a BootCaT-based corpus, where we used Acadian French words found within the other two corpora to query and scrape Google search results. Then, we compared these corpora to a known Acadian French corpus. Results can be found in Table 5.6. All three corpora exhibited traces of Acadian French. Our findings showed we can create an Acadian French corpus using previously proposed methods. Lastly, our third research question was *Is there one of these newly built corpora that surpasses others in quality and quantity?* To compare quantity, we look at the token size in Table 5.1. To look at quality, we look at how they compare to the known Acadian French corpus on all 22 measures of Acadian French. The known Acadian French corpus was created from the Acadian French dictionary's example sentences in each of its entries [22]. The results are shown in Table 5.6 and further summarized in Table 5.7. As we can see, no corpus excels in both quantity and quality. BootCaT has the highest number of tokens within its corpus, and r/acadie has the highest number of Acadian French properties found. However, BootCaT has the second highest number of Acadian French properties.

The contributions of the research conducted for this thesis are as follows:

1. We showed that NLP tools perform worse on Acadian French text than standard French.

2. We proposed a list of statistical corpus-based measures to indicate the presence of Acadian French within a given French corpus.

3. We showed that it is possible to create corpora containing Acadian French text using three different web-as-corpus methodologies

4. We showed that while all three had signs of Acadian French, BootCaT and social media scraping had the highest quantity and quality, respectively, of Acadian French text.

In this thesis, we tested two different current-day NLP tools on Acadian French. Future work could expand on this evaluation to include more NLP tools as additional testing could show further improvements an Acadian French corpus could bring to modern-day NLP tools. We also intended to confirm that these corpora could help off-the-shelf NLP tools perform better on Acadian French. Future work could show if training the models used with our corpora leads to better performance on Acadian French text. Additionally, we intended to use Twitter as the source of social media text since the location of origin of the tweets can be confirmed and queried. Since social media showed the highest number of measures of Acadian French, comparing platforms could perhaps improve the results even further. This also could shed light on which social media is best for location-based dialects, such as Acadian French.

An interesting line of work is if these results can be replicated with other French dialects, such as Quebec/Ontario French. If they can, this would not only further our understanding of Quebec/Ontario French but also create similar corpora that could be used to further compare these dialects to Acadian French. This could enable further corpus-based studies and lead to a better understanding of the nuances between the dialects.

From a linguistic perspective, another interesting line of work might be comparing past known Acadian French properties and see if they are still current in modern-day Acadian French text. Some of the measures appeared in more corpora than others, see Table 5.6, which could indicate that some are more present in the modern day than others.

With the constant evolution of language models, language models are learning more and more from low-resource languages, such as minority variants of languages [35]. We would like to see models able to identify, comprehend, and generate text of low-resource dialects, such as Acadian French. This would be useful in linguistic analysis of text and personalized text/speech recognition, as Acadian French is still a modern-day spoken dialect.

Bibliography

[1] Anne Abeillé, Lionel Clément, and Alexandra Kinyon, *Building a treebank for French*, Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00) (Athens, Greece) (M. Gavrilidou, G. Carayannis, S. Markantonatou, S. Piperidis, and G. Stainhauer, eds.), European Language Resources Association (ELRA), May 2000.

[2] Anne Abeillé, Lionel Clément, and Loïc Liégeois, *Un corpus arboré pour le français: le french treebank [a parsed corpus for french: the french treebank]*, Traitement Automatique des Langues **60** (2019), no. 2, 19–43.

[3] Rania Al-Sabbagh and Roxana Girju, *Yadac: Yet another dialectal Arabic corpus*, Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12) (Istanbul, Turkey), European Language Resources Association (ELRA), may 2012, pp. 2882–2889.

[4] Wedad Altawel, *La langue comme véhicule identitaire: analyse linguistique du brayonnaire (madawaska, nouveau-brunswick, canada)*, Universite de Moncton (Canada), 2021.

[5] Laurence Arrighi, *L'interrogation dans un corpus de français parlé en acadie. formes de la question et visées de l'interrogation*, Linx. Revue des linguistes de l'université Paris X Nanterre (2007), no. 57, 47–56.

[6] ———, *Le français parlé en acadie: description et construction d'une «variété»*, Minorités linguistiques et société (2014), no. 4, 100–125.

[7] Patricia Balcom, Louise Beaulieu, Gary R Butler, Wladyslaw Cichocki, and Ruth King, *The linguistic study of acadian french*, Canadian Journal of Linguistics/Revue canadienne de linguistique **53** (2008), no. 1, 1–5.

[8] Timothy Baldwin, Paul Cook, Marco Lui, Andrew MacKinlay, and Li Wang, *How noisy social media text, how diffrnt social media sources?*, Proceedings of the Sixth International Joint Conference on Natural Language Processing (Nagoya, Japan) (Ruslan Mitkov and Jong C. Park, eds.), Asian Federation of Natural Language Processing, October 2013, pp. 356–364.

[9] Marco Baroni and Silvia Bernardini, *BootCaT: Bootstrapping corpora and terms from the web*, Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04) (Lisbon, Portugal) (Maria Teresa Lino, Maria Francisca Xavier, Fátima Ferreira, Rute Costa, and Raquel Silva, eds.), European Language Resources Association (ELRA), May 2004.

[10] Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta, *The wacky wide web: a collection of very large linguistically processed web-crawled corpora*, Language resources and evaluation **43** (2009), 209–226.

[11] Tommy Berger, *Le chiac: entre langue des jeunes et langue des ancêtres : enjeux de nomination à travers les représentations linguistiques du chiac dans le sud-est du nouveau-brunswick*, (2020) (fra), Accepted: 2021-01-11T16:32:29Z.

[12] Henri Biahé, *Parlers hybrides en traduction: L'exemple du chiac et du cam-franglais*, Ph.D. thesis, 2017.

[13] Andreas Blombach, Natalie Dykes, Philipp Heinrich, Besim Kabashi, and Thomas Proisl, *A corpus of German Reddit exchanges (GeRedE)*, Proceed-

ings of the Twelfth Language Resources and Evaluation Conference (Marseille, France) (Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, eds.), European Language Resources Association, May 2020, pp. 6310–6316 (English).

[14] Peter F Brown, John Cocke, Stephen A Della Pietra, Vincent J Della Pietra, Frederick Jelinek, John Lafferty, Robert L Mercer, and Paul S Roossin, *A statistical approach to machine translation*, Computational linguistics **16** (1990), no. 2, 79–85.

[15] Henrico Brum and Maria das Graças Volpe Nunes, *Building a sentiment corpus of tweets in Brazilian Portuguese*, Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018) (Miyazaki, Japan), European Language Resources Association (ELRA), May 2018.

[16] Jamie Callan, Mark Hoy, Changkuk Yoo, and Le Zhao, *Clueweb09 data set*, 2009.

[17] Ian Campbell, *Chi-squared and fisher–irwin tests of two-by-two tables with small sample recommendations*, Statistics in medicine **26** (2007), no. 19, 3661–3675.

[18] Wladyslaw Cichocki, Sid-Ahmed Selouani, and Louise Beaulieu, *The racad speech corpus of new brunswick acadian french: design and applications*, Canadian Acoustics **36** (2008), no. 4, 3–10.

[19] William G Cochran, *Some methods for strengthening the common χ 2 tests*, Biometrics **10** (1954), no. 4, 417–451.

[20] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettle-

moyer, and Veselin Stoyanov, *Unsupervised cross-lingual representation learning at scale*, Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (Online), Association for Computational Linguistics, July 2020, pp. 8440–8451.

[21] Paul Cook and Laurel J. Brinton, *Building and evaluating web corpora representing national varieties of english*, Language Resources and Evaluation **51** (2017), no. 3, 643–662.

[22] Yves Cormier and Russon Wooldridge, *Dictionnaire du francais acadien*, University of Toronto Quarterly **70** (2000), no. 1, 172.

[23] Mark Davies and Robert Fuchs, *Expanding horizons in the study of world englishes with the 1.9 billion word global web-based english corpus (glowbe)*, English World-Wide **36** (2015), no. 1, 1–28.

[24] Guy De Pauw, *Developing Linguistic Corpora—A Guide to Good Practice*, Literary and Linguistic Computing **22** (2006), no. 1, 101–102.

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

[26] Jonathan Dunn, *Mapping languages: The corpus of global language use*, Language Resources and Evaluation **54** (2020), 999–1018.

[27] Andreas Eisele and Yu Chen, *MultiUN: A multilingual corpus from united nation documents*, Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10) (Valletta, Malta) (Nicoletta Calzolari,

Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, eds.), European Language Resources Association (ELRA), May 2010.

[28] Adriano Ferraresi, Eros Zanchetta, Marco Baroni, and Silvia Bernardini, *Introducing and evaluating ukwac, a very large web-derived corpus of english*, Proceedings of the 4th Web as Corpus Workshop (WAC-4) Can we beat Google, 2008, p. 47–54.

[29] W. N. Francis and H. Kucera, *Brown corpus manual*, Tech. report, Department of Linguistics, Brown University, Providence, Rhode Island, US, 1979.

[30] Karine Gauvin, *Bdlp·acadie*, 2004, Last accessed 24 février 2023.

[31] Barry M Gough, *Historical dictionary of canada*, Scarecrow Press, 2010.

[32] Sidney Greenbaum and Gerald Nelson, *The international corpus of english (ice) project*, World Englishes **15** (1996), no. 1, 3–15.

[33] Naomi ES Griffiths, *Contexts of acadian history, 1686-1784*, McGill-Queen's Press-MQUP, 1992.

[34] Emiliano Raul Guevara, *NoWaC: a large web-based corpus for Norwegian*, Proceedings of the NAACL HLT 2010 Sixth Web as Corpus Workshop (NAACL-HLT, Los Angeles) (Adam Kilgarriff and Dekang Lin, eds.), Association for Computational Linguistics, June 2010, pp. 1–7.

[35] Michael A. Hedderich, Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow, *A survey on recent approaches for natural language processing in low-resource scenarios*, Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Online), Association for Computational Linguistics, June 2021, pp. 2545–2568.

[36] Maurice Holder, Anne Macies, and Rolf Turner, *La diphthongue «oi» dans le parler «brayon» d'edmundston, nouveau-brunswick*, Linguistica Atlantica (1992), 17–54.

[37] Mikhail Khodak, Nikunj Saunshi, and Kiran Vodrahalli, *A large self-annotated corpus for sarcasm*, Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018) (Miyazaki, Japan), European Language Resources Association (ELRA), May 2018.

[38] Adam Kilgarriff, *Simple maths for keywords*, Proc. Corpus Linguistics, vol. 6, 2009.

[39] Ruth Elizabeth King, *Acadian french in time and space: A study in morphosyntax and comparative sociolinguistics*, (No Title) (2013).

[40] Philipp Koehn, *Europarl: A parallel corpus for statistical machine translation*, Proceedings of Machine Translation Summit X: Papers (Phuket, Thailand), September 13-15 2005, pp. 79–86.

[41] H. Kucera and W. N. Francis, *Computational analysis of present-day american english*, Brown University Press, Providence, RI, 1967.

[42] Geoffrey Leech et al., *100 million words of english: the british national corpus (bnc)*, Language research **28** (1992), no. 1, 1–13.

[43] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov, *Roberta: A robustly optimized BERT pretraining approach*, CoRR **abs/1907.11692** (2019).

[44] Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky, *The Stanford CoreNLP natural language processing toolkit*, Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations (Baltimore, Maryland) (Kalina

Bontcheva and Jingbo Zhu, eds.), Association for Computational Linguistics, June 2014, pp. 55–60.

[45] France Martineau, *Corpus fran corpus du français d'amérique du nord, élaboré dans le cadre du projet le français à la mesure d'un continent: un patrimoine en partage*, 2011.

[46] France Martineau and Marie-Claude Séguin, *Le corpus fran: réseaux et maillages en amérique française*, Corpus (2016), no. 15.

[47] Pierre André Ménard and Desislava Aleksandrova, *A French corpus of Québec's parliamentary debates*, Proceedings of the Workshop ParlaCLARIN III within the 13th Language Resources and Evaluation Conference (Marseille, France), European Language Resources Association, June 2022, pp. 25–32.

[48] Aaron Mueller, Garrett Nicolai, Panayiota Petrou-Zeniou, Natalia Talmina, and Tal Linzen, *Cross-linguistic syntactic evaluation of word prediction models*, Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (Online) (Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, eds.), Association for Computational Linguistics, July 2020, pp. 5523–5539.

[49] Brian Murphy and Egon W. Stemle, *PaddyWaC: A minimally-supervised web-corpus of hiberno-English*, Proceedings of the First Workshop on Algorithms and Resources for Modelling of Dialects and Language Varieties (Edinburgh, Scotland) (Jeremy Jancsary, Friedrich Neubarth, and Harald Trost, eds.), Association for Computational Linguistics, July 2011, pp. 22–29.

[50] Marie-Ève Perrot, *Aspects fondamentaux du métissage français/anglais dans le chiac de moncton (nouveau-brunswick, canada)*, Ph.D. thesis, Paris 3, 1995.

[51] ———, *Tu worries about ça, toi? métissage et restructurations dans le chiac de moncton*, Linx **33** (1995), no. 2, 79–85.

[52] Marie-Ève. Perrot, *Le trajet linguistique des emprunts dans le chiac de moncton : quelques observations*, Minorités linguistiques et société / Linguistic Minorities and Society (2014), no. 4, 200–218 (fr).

[53] Marie-Ève Perrot, *Comparer les emprunts à l'anglais dans les variétés de français acadien: méthodes et enjeux*, Arrighi L. et Gauvin K (2018), 113–130.

[54] Jan Pomikálek, *Removing boilerplate and duplicate content from web corpora*, Ph.D. thesis, Masaryk university, Faculty of informatics, Brno, Czech Republic, 2011.

[55] Louise Péronnet, « *lexique d'acadianismes* », *dans le dictionnaire en ligne usito.*, 2013, Last accessed 24 février 2023 (version 1676668985).

[56] Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning, *Stanza: A python natural language processing toolkit for many human languages*, Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations (Online) (Asli Celikyilmaz and Tsung-Hsien Wen, eds.), Association for Computational Linguistics, July 2020, pp. 101–108.

[57] Randolph Quirk, *Language varieties and standard language*, English today **6** (1990), no. 1, 3–10.

[58] Roland Schäfer and Felix Bildhauer, *Web corpus construction*, Synthesis Lectures on Human Language Technologies **6** (2013), no. 4, 1–145.

[59] Charlene Soucy-Godby and Gert Michaud, *Brayonnaire: Petit dictionnaire brayon / français / anglais*, (2014).

[60] ———, *Brayonnaire – partie 2: Petit dictionnaire brayon / français / anglais*, (2016).

[61] Johansson. Stig, Geoffrey N. Leech, and Helen Goodluck, *Manual of information to accompany the lancaster-oslo : Bergen corpus of british english, for use with digital computers*, Department of English, University of Oslo, 1978.

[62] Ronald J Tallarida, Rodney B Murray, Ronald J Tallarida, and Rodney B Murray, *Chi-square test*, Manual of pharmacologic calculations: With computer programs (1987), 140–142.

[63] Spencer Trerice, *Entre fierté et mépris: le rapport ambivalent à l'égard du chiac dans" pour sûr" de france daigle*, Ph.D. thesis, 2016.

[64] Motoko Ueyama, Marco Baroni, et al., *Automated construction and evaluation of japanese web-based reference corpora*, Proceedings of Corpus Linguistics 2005 (2005).

[65] Graham J. G. Upton, *A comparison of alternative tests for the 2 × 2 comparative trial*, Journal of the Royal Statistical Society. Series A (General) **145** (1982), no. 1, 86–105.

[66] Graham JG Upton, *Fisher's exact test*, Journal of the Royal Statistical Society: Series A (Statistics in Society) **155** (1992), no. 3, 395–402.

[67] Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave, *CCNet: Extracting high quality monolingual datasets from web crawl data*, Proceedings of the Twelfth Language Resources and Evaluation Conference (Marseille, France), European Language Resources Association, May 2020, pp. 4003–4012 (English).

[68] Raphaële Wiesmath, *Le français acadien: Analyse syntaxique d'un corpus oral recueilli au nouveau-brunswick/canada*, Le français acadien (2006), 1–278.

Vita

Candidate's full name: Jérémy Zacharie Robichaud
University attended (with dates and degrees obtained): Bachelor of Computer Science University of New Brunswick, 2022
Publications: N/A
Conference Presentations: N/A