WaCadie: Towards an Acadian French Corpus

Jérémy Robichaud, Paul Cook
Faculty of Computer Science, University of New Brunswick, Canada
jrobich9@unb.ca, paul.cook@unb.ca

Abstract

Corpora are important assets within the natural language processing (NLP) and linguistics communities, as they allow the training of models and corpus-based studies of languages. However, corpora do not exist for many languages and language varieties, such as Acadian French. In this paper, we first show that off-the-shelf NLP systems perform more poorly on Acadian French than on standard French. An Acadian French corpus could, therefore, potentially be used to improve NLP models for this dialect. Then, leveraging web-as-corpus methodologies, specifically BootCaT, domain crawling, and social media scraping, we create three corpora of Acadian French. To evaluate these corpora, drawing on the linguistic literature on Acadian French, we propose 22 statistical corpus-based measures of the extent to which a corpus is Acadian French. We use these measures to compare these newly built corpora to known Acadian French text and find that all three corpora include Acadian French content.

Keywords: Web Corpora, Low-Resource Languages, Acadian French

1. Introduction

Corpora are important resources within the natural language processing (NLP) and linguistics communities as they allow the training of NLP models and corpus-based studies of languages (Kucera and Francis, 1967; Francis and Kucera, 1979; Leech et al., 1992; Greenbaum and Nelson, 1996). Due to their importance, various methodologies have been developed to create corpora. In the last twenty years, many corpora have been created by leveraging the internet as an information bank (Baroni and Bernardini, 2004; Cook and Brinton, 2017; Wenzek et al., 2020; Dunn, 2020). This corpus-building process is known as web-as-corpus and has been used to create corpora of many languages (Schäfer and Bildhauer, 2013).

There exist multiple ways of using the internet to create a corpus (Schäfer and Bildhauer, 2013). Baroni and Bernardini (2004) used search engines (e.g., Google) to query URLs based on words related to the targeted corpus. This method became known as BootCaT. Another method is creating corpora by leveraging the domains of URLs on the internet. The last portion of a URL (e.g., *.ca*) is the domain, and it represents a certain group to which URLs belong (e.g., *.ca* means the Canadian domain). We can use these URL structures to gather websites within a certain group and extract texts from them (Cook and Brinton, 2017). This is called domain crawling. We can further use snapshots of the web, such as Common Crawl (Wenzek et al., 2020), to acquire large amounts of web content from domains of interest (Dunn, 2020). Lastly, we can use social media to create a corpus by extracting the text of user posts (e.g., Baldwin et al., 2013). This has become known as social media scraping.

Acadians are a French minority group of people that can be found throughout the predominantly English-speaking Atlantic Provinces of Canada and Maine and Louisiana in the United States of America. (In Louisiana, this group is known as *cajun* (Griffiths, 1992)). However, the largest portion of Acadians resides in the province of New Brunswick, Canada (Gough, 2010; Arrighi, 2014; Wiesmath, 2006). Most Acadians speak a variety of French known as Acadian French (Balcom et al., 2008). Additionally, multiple varieties of Acadian French exist, such as *Chiac* and *Brayon* in the south-east and north-west regions of New Brunswick, respectively. To our knowledge, little research has been done on the presence of Acadian French on the web. There currently does not exist a large written corpus of Acadian French, but there do exist a few smaller spoken corpora (Wiesmath, 2006; Arrighi, 2007; Berger, 2020; Perrot, 1995a).

After discussing related work (Section 2), this paper evaluates the performance of off-the-shelf NLP tools on Acadian French (Section 3). Specifically, we examine systems for two tasks: masked word prediction and part-of-speech tagging. We find that these systems perform worse on Acadian French than standard French. This indicates a need for a corpus of Acadian French, which could be used for training NLP systems to improve their performance on Acadian French.

We then create three Acadian French web corpora using different web-as-corpus methodologies: BootCaT, domain crawling, and social media scraping (Section 4). It is, however, unclear whether these corpora indeed include Acadian French.

In Section 5, we evaluate the corpora for whether they include Acadian French. Drawing on linguistic literature on Acadian French, we propose 22 statistical measures that indicate that a corpus includes Acadian French. We evaluate these measures and show that they collectively identify Acadian French

LREC-COLING 2024, pages 17421–17432

text. We then apply these measures to the newly constructed corpora and find evidence that all three corpora include Acadian French. In particular, our social media corpus has the highest number of Acadian French characteristics but also the smallest size. The BootCaT corpus, on the other hand, was similar in terms of these characteristics to an Acadian French reference corpus, but two orders of magnitude larger than the social media corpus.

2. Related Works

Over the last roughly 60 years, a variety of research has created corpora and showcased their value (Kucera and Francis, 1967; Francis and Kucera, 1979; Stig et al., 1978; Brown et al., 1990; Leech et al., 1992; Greenbaum and Nelson, 1996; Abeillé et al., 2000; Koehn, 2005; Cichocki et al., 2008; Eisele and Chen, 2010; Martineau, 2011; Martineau and Séguin, 2016; Ménard and Aleksandrova, 2022). In the last two decades, using the web as a source of text for corpora has become common (Baroni and Bernardini, 2004; Schäfer and Bildhauer, 2013; Baldwin et al., 2013), with multiple methods of using the web to build corpora being proposed (Schäfer and Bildhauer, 2013). Here we will consider BootCaT, domain crawling, and social media scraping.

An early approach to web corpus construction is BootCaT, which uses commercial search engines (e.g. Google or Yahoo) to query websites that may hold valuable text for the intended corpus (Baroni and Bernardini, 2004). 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 querying. However, since search engines are black box tools, the BootCaT method gives little control over which URLs are returned. It has nevertheless been found to be an effective method for acquiring topic-specific text (e.g., psychiatric articles in both English and Italian) (Baroni and Bernardini, 2004). BootCaT has also been used to create corpora for a range of languages (Ueyama et al., 2005; Ferraresi et al., 2008). Additionally, BootCaT has been used to build corpora for regional and minority variants of languages and languages with little internet presence (Guevara, 2010; Murphy and Stemle, 2011; Davies and Fuchs, 2015).

Domain crawling leverages the URLs of websites and collects the data of all websites crawled within a specific domain (Schäfer and Bildhauer, 2013). This method is often utilized for region-specific domains, such as .ca or .uk. Cook and Brinton (2017) used domain crawling to show that national top-level domains (e.g., *.ca* and *.uk*) yield content that reflects their corresponding variety of English. They used ClueWeb09 (Callan et al., 2009), a 2009 snapshot of crawled websites filtered by domain and language, to create corpora of national varieties of English. Dunn (2020) builds corpora using multiple domains for countries using the Common Crawl database (Wenzek et al., 2020).

Social media scraping uses social media posts as a source of text. Baldwin et al. (2013) address some concerns about social media corpora and conclude that noise in social media text can be dealt with using NLP tools and that differences in grammatical structure between social media and standard edited text are relatively small. Al-Sabbagh and Girju (2012) and Brum and Volpe Nunes (2018) create Twitter-based corpora representing minority language varieties, Egyptian Arabic and Brazilian Portuguese, respectively. Khodak et al. (2018) shows the possibilities of using Reddit to create a corpus focusing on sarcasm. Blombach et al. (2020) created a German corpus of 270M tokens from Reddit.

Most of the Acadian population resides in New Brunswick (Arrighi, 2014; Wiesmath, 2006). Acadian French has unique features not seen in standard French (Cichocki et al., 2008; Berger, 2020; Perrot, 2014, 2018; Altawel, 2021). Additionally, variants of Acadian French exist in New Brunswick. Brayon, a variety of Acadian French spoken in the north-west, borrows from neighbouring French-speaking Quebec (Altawel, 2021). Brayon has unique features not seen elsewhere in New Brunswick (Holder et al., 1992), such as using adverbs with the suffix *-eux* more frequently (Altawel, 2021) and having unique expressions (Soucy-Godby and Michaud, 2014, 2016). Chiac, spoken in the south-east, borrows from English (Berger, 2020). As such, Chiac also has unique features, such as code-mixing, not seen in other varieties (Perrot, 1995b, 2014; Berger, 2020). Brayon and Chiac are both commonly spoken varieties of Acadian French and, as such, are targeted equally in our research. We will use properties of Acadian French, including Brayon and Chiac, to develop measures of the extent to which a corpus exhibits Acadian French properties and could, therefore, be considered to be Acadian French.

3. Preformance of NLP Tools on Acadian French

We evaluate the need for Acadian French corpora in the current-day NLP space by comparing the performance of off-the-shelf NLP tools when used on standard French and Acadian French. If these tools perform more poorly on Acadian French, then an Acadian French corpus would be beneficial, and could potentially be leveraged to train NLP systems for Acadian French.

3.1. Known French Corpora

We require Acadian French and non-Acadian French corpora for this comparison. These corpora must be of these varieties of French and also otherwise comparable, for example, with respect to text types. We address this by creating corpora from example sentences found within dictionaries of the target French varieties. The first corpus uses all example sentences from the Acadian French dictionary (Cormier and Wooldridge, 2000). The second corpus is from Wiktionnaire, a French online dictionary.[1] We extract example sentences from Wiktionnaire for all headwords that are also listed in GNU Aspell's list of French words.[2] We refer to this corpus as WiktionaryFR. We apply exact deduplication, near-deduplication, and tokenization (Section 4.2) to both corpora. AcadianDictionary totalled 69k tokens and 4,009 sentences, while WiktionaryFR totalled 2,6M tokens and 96k sentences.

3.2. Masked Language Modelling

In our first test, we mask a token in a given sentence and calculate the accuracy at which a language model can predict the correct token and the perplexity of the predictions (Mueller et al., 2020). We consider XLM-RoBERTa (Conneau et al., 2020), a multilingual RoBERTa model trained on 100 languages. We use all sentences in AcadianDictionary for this analysis and randomly sample the same number of sentences (4,009) from WiktionaryFR. This ensures we compare similar size corpora for this analysis. We iterate through each token of each sentence within a corpus, masking each token in turn, and examine the model's prediction for that token. We compute accuracy-at-k, for k=1, 5, and 10, in which the model is scored as correct if the target token is within the top k predictions. We use the chi-squared test (Tallarida et al., 1987) to determine whether differences in accuracy-at-k are significant. We also compute the perplexity of the model's predictions.[3]

Results are shown in Table 1. XLM-RoBERTa performs better on standard French than on Acadian French in terms of accuracy-at-k, for each value of k considered, and perplexity. The differences for accuracy-at-k are significant ($p < 0.05$).

3.3. POS Tagging

In our second test, we consider POS tagging. We tag both AcadianDictionary and WiktionaryFR using the StanfordNLP pipeline (Qi et al., 2020). This tagger tags words that cannot be assigned a real POS, such as unknown words, with the tag 'X'.[4] In our analysis, we consider the number of tokens tagged X, as well as the number of sentences with at least one token tagged X, in each corpus. We again compare differences between corpora using chi-squared tests.

Results are shown in Table 2. We note that differences in terms of both tokens and sentences are significant. A higher percentage of tokens are tagged X for Acadian French than for standard French. This is as expected because the POS tagger has not specifically been trained on Acadian French text. However, a higher percentage of sentences in the standard French corpus include at least one token tagged X. Analyzing the most common tokens tagged X in WiktionaryFR, we see tokens that indicate listed examples (e.g., *i*, *ii*, *iii*, *iv*), which would not be expected to co-occur in a sentence. Indeed, WiktionaryFR averages 4.48 tokens tagged X per sentence containing at least one token tagged X, whereas this is 5.8 for AcadianDictionary.

These findings indicate that, at a minimum, Acadian French text is harder to process at the token level for the part-of-speech tagger. These POS tagging findings, and our previous findings for masked language modelling, indicate that off-the-shelf NLP tools perform worse on Acadian French than standard French text. This suggests that Acadian French text is likely not seen in the training data for these tools. Creating an Acadian French corpus could help to alleviate this performance disparity by enabling NLP systems to be trained for Acadian French.

4. Corpus Construction

In this section, we discuss the construction of Acadian French web corpora using domain crawling, social media scraping, and BootCaT and the post-processing pipeline applied to each corpus.

4.1. Data Retreival

Our first corpus uses French New Brunswick domain (i.e., *.nb.ca*) websites in the Common Crawl database (Wenzek et al., 2020). We refer to this corpus as *CC_NB_Domain*. We focus on *.nb.ca* because New Brunswick has the largest Acadian population per capita. Common Crawl's data is publicly available through its index server.[5] We queried all *.nb.ca* websites. Common Crawl adds metadata to each download, including the language identified within the website by its internal language

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

[2] http://aspell.net/

[3] The chi-squared test is not directly applicable to the differences in perplexity, and so we do not compute it for this measure.

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

[5] https://index.commoncrawl.org/

Corpus	Accuracy-at-1	Accuracy-at-5	Accuracy-at-10	Perplexity
AcadianDictionary	0.382	0.546	0.589	8.129
WiktionaryFR	0.419	0.586	0.629	7.935
Chi-Squared p	6.632e-51	5.755e-59	5.033-62	

Table 1: Masked language modelling results.

Corpus	Num. tokens	Num. sentences
AcadianDictionary	464 (0.670%)	80 (1.996%)
WiktionaryFR	12,790 (0.487%)	2855 (2.968%)
Chi-Squared p	4.695e-10	0.005

Table 2: The number of tokens tagged X, and the number of sentences with at least one token tagged X, in each corpus. The corresponding percentage of tokens and sentences is shown in parentheses.

identifier. (We also apply language identification in our post-processing pipeline (Section 4.2).) This process allows us to query websites identified as containing French text and we obtain 2324 URLs, including websites that are exclusively French and a mix of French and another language. We download the corresponding HTML from those URLs directly from their data server.[6] The HTML encoding stored within the server is not always correctly indicated; we address this in our post-processing pipeline.

Our second corpus is a social media-based corpus. We use Reddit as it offers accessible APIs to collect data while the data is topically organized through subreddits. We use the subreddit *r/acadie*, which focuses on Acadian themes. Reddit's data can be harvested through a Python wrapper called PRAW.[7] PRAW allows us to search for a specific subreddit and gets all its posts and each comment within each post. 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 Common Crawl corpus. We refer to this corpus as r/acadie.

Our last corpus is our BootCaT corpus. We issue queries containing 3-tuples of French words, which include Acadian terms, to a search engine to retrieve documents containing Acadian French. However, when multiple Acadian words are included in the search, we observed that the results tend to be dominated by online Acadian dictionaries, as opposed to Acadian French text. To combat this, and avoid retrieving online dictionaries, we combine one Acadian French word (the target word) with two medium-frequency French words from the Aspell dictionary, and mandate that the target word must occur within the results. To avoid using very rare Acadian terms in our queries, which might give very few search results, we only use Acadian terms found in our other corpora (i.e., CC_NB_Domain and r/acadie). We randomly generate 45 tuples and restrict the search to only HTML documents. We use the BootCaT front-end to process our tuples and scrape the returned URLs.[8] We choose a small amount of tuples as this is sufficient to explore the viability of this method for creating an Acadian corpus, while staying within the terms of service. Up to 10 documents were returned per tuple.

4.2. Corpus Processing Pipeline

For each of our corpora, we apply the same processing pipeline to create a corpus from HTML files. One challenge is that HTML files may have different encodings (e.g., UTF-8, Latin-1). We chose UTF-8 as the project-wide encoding standard. We use the Chardet Python library to guess the original encoding of each HTML file and re-encode it to UTF-8.[9]

To extract the text from HTML documents, we use Justext (Pomikálek, 2011) as it allows us to get the text while filtering out boilerplate content. It takes in HTML and outputs clusters of paragraphs of text. Paragraphs are flagged as boilerplate if they are within the documented HTML boilerplate patterns and are removed.

We then verify that the text is French using langid.py[10], a Python language identifier. We run langid.py at the document level and reject all documents not classified as French. Even though Acadian French, specifically Chiac, includes many English words, it contains primarily French words and thus would be likely flagged as French by langid.py, which tends to classify multilingual documents according to the predominant language they include.

Different UTF-8 characters could represent two identical-seeming accents. For example, *è* could be represented by its Latin letter (U+00E8) or by combining the letter *e* and a grave accent ` (U+0060). Because of this, we normalize the accents using Unicodedata.[11] Unicode normaliza-

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

[7] https://praw.readthedocs.io/en/stable/

[8] https://bootcat.dipintra.it/

[9] https://pypi.org/project/chardet/

[10] https://github.com/saffsd/langid.py

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

Corpus	Tokens	Types	Sent.	Docs
BootCaT	1,208,629	66,555	74,419	240
CC_NB_Domain	376,668	23,505	20,009	2,324
r/acadie	56,258	12,790	9,907	801

Table 3: The number of tokens, types, sentences, and documents in each web corpus built.

tion handles two types of normalization: compatibility equivalence and canonical equivalence. Compatibility equivalence normalizes the stylistic aspects of the text (e.g. font, subscript, fractions). Canonical equivalence normalizes the characters. This means it normalizes combined character sequences that form another character. We apply both types of normalization to the text in our corpora. We apply case folding to all corpora.

Repeated content in a corpus can skew analyses and thus should be removed. We hash all paragraph objects and scan for duplicates in the corpus. If we find an identical hashing, we remove its most recently seen copy. This ensures the corpus contains no duplicate paragraphs.

Similarly, we do not want near-identical content, which gives very little new information in a corpus. To remove near-duplicate content, we use PyOnion,[12] a Python implementation of the Onion corpus deduplication tool (Pomikálek, 2011). It iterates through the paragraphs in a corpus and compares the n-grams in each paragraph with the set of n-grams observed so far. Any paragraph above a set threshold for n-gram overlap is removed. We used 5-grams (the default setting for Onion) with a threshold of 0.25, which Pomikálek notes balances removing duplicate content and losing corpus data.

Lastly, we ready the corpora for analysis. We use Stanza (Qi et al., 2020), a Python library that pipelines the Java Stanford CoreNLP.[13] We use Stanza's tokenization for one final filtering of very long sentences (250+ words) and English-only sentences since those are most likely quotes or titles of articles. The remaining data is then re-processed by the Stanza pipeline for part-of-speech tagging, lemmatization, and dependency parsing.

Table 3 contains the number of tokens, types, sentences, and documents in each newly built corpus.

5. Corpus Analysis

In this section, we discuss our analysis of the newly built corpora. This includes keyword analysis and comparing corpora based on Acadian characteristics found within them.

[12] https://pypi.org/project/pyonion/

[13] https://stanfordnlp.github.io/CoreNLP/

5.1. French Web Corpus

For our analyses, we require a web corpus known to be of non-Acadian French. We use frWaC (Baroni et al., 2009), a French web corpus of roughly 1.8B tokens built by scraping *.fr* domain websites. FrWaC is publicly available.[14] In addition to the full corpus, smaller samples are available. We use the 10M token sample, which is larger than any of our Acadian corpora.

5.2. Keyword Analysis

We first analyze our Acadian corpora by comparing their keywords with respect to frWaC. We use the approach of Kilgarriff (2009) for computing keywords. The keywordness score (KW) for a word w is calculated as shown below:

$$\mathrm{KW}(w) = \frac{\mathrm{fpm}_{\mathrm{fc}}(w) + c}{\mathrm{fpm}_{\mathrm{rc}}(w) + c} \quad (1)$$

where $\mathrm{fpm}_x(w)$ is the frequency per million of word w in corpus x; fc is the focus corpus, which will be one of the Acadian corpora from Section 4; rc is the reference corpus, which is frWaC; and c is a constant which we set to 100 following Kilgarriff et al. (2010). We consider the top-20 keywords for each corpus.

BootCaT has the following top-20 keywords: *u, r, e, o, t, i, québec, n, canada, autochtones, d, québécois, yolanda, caribou, p, steven, -là, l'on, canadiens, q*. The keywords *québec*, *canada*, *autochtones* ('indigenous'), *québécois*, and *canadiens* indicate that the corpus includes many documents about Canada and peoples within Canada, with a focus on Québec. Keywords such as *caribou* also suggest a topical focus on Canada. There are ten single-character keywords. These represent the initials of speakers from recorded discussions. *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 top-20 keywords are: *nouveau-brunswick, officielles, élèves, canada, moncton, langues, école, élève, province, scolaire, commissaire, -vous, linguistiques, francophone, coucher, fredericton, langue, salle, soins, assurance*. The corpus has a focus toward the education sector of New Brunswick as indicated by the keywords *élèves* ('students'), *école* ('school'), *élève* ('student'), *scolaire* ('scholar'), and *salle* ('classroom'). This is because the education sector of the province of New Brunswick uses the domain *nbed.nb.ca*. Additionally, in the past, New Brunswick municipal sites were hosted within the *.nb.ca* domain. This reflects keywords related to

[14] https://wacky.sslmit.unibo.it/

cities and residences such as *moncton*, *coucher* ('sleep'), and *fredericton*. Documents from the domain languesofficielles.nb.ca were prevalent within the corpus, which is reflected in keywords such as *officielles* ('official'), *langues* ('languages'), *commissaire* ('commissioner'), *linguistiques* ('linguistics'), *francophone*, and *langue* ('language'). Again there are keywords related to Canada, and in this case New Brunswick, such as *nouveau-brunswick*, *canada*, and *province*. The keywords *soins* ('care') and *assurance* ('insurance') are seen mostly in a health insurance document. And lastly, *-vous* is a word seen in standard French.

r/acadie has the following top-20 keywords: *acadien, acadie, acadiens, **, acadienne, chiac, pis, *, francophones, québec, the, mot, nouveau-brunswick, francophonie, and, canada, to, acadiennes, québécois, langue*. r/acadie has a topical Acadian focus, with the keywords *acadien*, *acadie*, *acadiens*, *acadienne*, *chiac*, and *acadiennes* all being directly related to Acadians and Acadian French. Acadian French dialect words, including English words commonly used in Chiac, appear as well with *pis*, *the*, *and*, and *to*. *francophones*, *francophonie*, and *langue* directly relate to French and language. *québec*, *nouveau-brunswick*, *canada*, and *québécois* all indicate the corpus includes Canadian topics. *mot* (*word*) was observed in posts discussing words and the context in which you could find them. Lastly, ** and * are stylistic indicators found within Reddit text.[15]

These findings suggest that all three corpora are oriented toward Canada and French in their content. CC_NB_Domain contains New Brunswick content, r/acadie contains Acadian content, and BootCaT seems to lean towards Québec content.

5.3. Measures of Acadian French

To measure the extent to which a corpus is Acadian French, we create 22 statistical corpus-based measures drawing on previously-noted properties of Acadian French (Perrot, 1995a; Cormier and Wooldridge, 2000; Wiesmath, 2006; King, 2013; Perrot, 2014; Trerice, 2016; Biahé, 2017; Perrot, 2018; Berger, 2020; Altawel, 2021; Soucy-Godby and Michaud, 2014, 2016). Table 3 summarizes the measures. The measures tap into general properties of Acadian French, as well as properties specific to Brayon and Chiac, as noted in the Table. Each measure is formulated to be high if a corpus is Acadian and low otherwise.

We calculate these measures for each corpus. By comparing two corpora with respect to these measures, we can determine whether one corpus is more Acadian than the other.

	Acadian Properties	Formula
Acadian	Acadian French Tokens	# Acadian French tokens / N
	Acadian French Types	# Acadian French types / V
	Auxiliary *Avoir* Ratio	# auxiliary *avoir* tokens / # auxiliary tokens
	Acadian Conjunctions	# Acadian conjunction tokens / # conjunction tokens
	Infinitive verb with prep.	# Acadian preposition + inf. verb tokens / # inf. verb tokens
	Questions Containing *ti*	# questions contaning the token *ti* / # questions
	English verb tokens	# English verb tokens / # verb tokens
	English verb types	# English verb types / # verb types
	Point negation	# *point* adverb tokens / # *point* + *pas* adverb tokens
Brayon	Brayon Tokens	# Brayon French tokens / N
	Brayon Types	# Brayon French types / V
	Brayon Expressions	# sentences with a Brayon expression / # sentences
	Adverbs ending in *-eux*	# adverb tokens with the suffix *-eux* / # adverb tokens
Chiac	English Tokens	# English tokens / N
	English Types	# English types / V
	3rd Pers. Pl. *-ont* Verbs	# third person plural *-ont* verb tokens / # third person plural verb tokens
	-ly Adverb Tokens	# *-ly* suffixed English adv. tokens / # *-ly* English + *-ment* French adv. tokens
	-ly Adverb Types	# *-ly* suffixed English adv. types / # *-ly* English + *-ment* French adv. types
	Instances of *you know*	# *you know* tokens / # *you know* tokens + French alternatives
	Instances of *right*	# adverbial *right* tokens / # abverb tokens
	Instances of *back*	# adverbial *back* tokens / # abverb tokens
	Instances of *own*	# *own* tokens / # *own* tokens + French alternatives

Table 4: Formulas for measures of Acadian French.

5.4. Benchmark Comparison

Before we use these measures of Acadian French to determine whether the Acadian corpora we constructed are indeed Acadian French, we first evaluate these measures to determine whether they do, in fact, behave as expected.

We compare AcadianDictionary, a known Acadian French corpus, to WiktionaryFR, a comparable corpus known to be standard French (Section 3.1). If the measures are higher for AcadianDictionary than for WiktionaryFR, then the measures behave as expected, and correctly indicate that Acadian French is more Acadian than standard French.

We apply Fisher's Exact Test to determine whether differences between the corpora are significant. Fisher's Exact Test, while exact for small samples, is known for being conservative at a larger scale (Upton, 1982). To apply Fisher's Exact Test, we transform our measures into 2x2 contingency matrices. To do this, we count the number of times the measure is seen in each corpus (i.e. the numerator of the measure from Table 4) and the number

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

	Acadian Tokens	non-Acadian Tokens
AcadianDict.	1,808	67,426
WiktionaryFR	21,270	2,606,092

Table 5: Example 2x2 matrix used to calculate Fisher's Exact Test for the measure Acadian French Tokens We count the number of Acadian, and non-Acadian, tokens in each corpus.

	Acadian French Measures	AcadianDictionary
Acadian	Acadian French Tokens	
	Acadian French Types	
	Auxiliary *Avoir* Ratio	
	Acadian Conjunctions	
	Acadian Relative Pronouns	
	Questions Containing *ti*	
	English Verb Tokens	
	English 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	
	Third Person Plural *-ont* Verbs	
	-ly Adverb Tokens	
	-ly Adverb Types	
	Instances of *you know*	
	Instances of *right*	
	Instances of *back*	
	Instances of *own*	

Table 6: Comparison of the target corpus AcadianDictionary to the reference corpus WiktionaryFR.

of times it is not seen (i.e. denominator − numerator). Table 5 shows an example 2x2 matrix for the Acadian French Tokens measure.

Table 5 shows the results of these comparisons. Measures that are significantly different ($p < 0.05$) between the two corpora are shown in green and red. Green indicates that the measure was higher in AcadianDictionary (the target corpus) and red indicates that the measure was higher in the reference corpus. Yellow indicates that the difference is not significant.

The Acadian French measures are, in most cases, higher for AcadianDictionary than for WiktionaryFR, or there is no difference between the corpora, with two exceptions: Acadian relative pronouns and English types. This indicates that the measures overall behave as expected, and are able to correctly recognize Acadian French as more Acadian than standard French. Future work could further examine the measures that do not behave as expected to better understand why this is the case.

5.5. Results

In this section, we apply the measures of Acadian French to the Acadian corpora we built — BootCaT, CC_NB_Domain, and r/acadie — to determine whether these corpora are indeed Acadian. We compare each Acadian corpus (the target corpus) to the same reference corpus: frWaC. We consider an additional target corpus, AcadianDictionary, as a point of comparison. We compare target and reference corpora in the same way as in Section 5.4.

Table 6 shows the results of these comparisons. Again, measures that are significantly different ($p < 0.05$) between the target and reference corpora are shown in green and red. Green indicates that the measure was higher in the target corpus, and red indicates that the measure was higher in the reference corpus (frWaC). Yellow indicates that the difference is not significant.

The measures of Acadian French Tokens and Acadian French Types are higher for all target corpora than the reference corpus. These measures focus on Acadian-specific vocabulary and capture the total number of usages of these words (Acadian French Tokens) as well as the diversity of these words in a corpus (Acadian French Types). All three corpora we built contain more Acadian French tokens and types than the reference corpus, indicating that each corpus we built does indeed include Acadian French. We now further consider the findings for each corpus.

BootCaT shows identical results for Brayon measures as AcadianDictionary. General Acadian French measures are also very similar to AcadianDictionary except for English Verb Tokens. Most Chiac measures are not higher for this corpus than the reference corpus. This corpus has the highest amount of significant Brayon French measures of the corpora we built. This could be due to the overlap between Brayon and Québec French. Québec, a larger province in Canada, potentially has more internet presence, and thus, Québec Frech could be more prominent in search engine results. This influence of Québec was also observed in the keyword analysis. *Québec* and *québécois* were found to be keywords for this corpus.

For Common Crawl, most measures are not significant. This is likely due to size and content. This corpus was restricted to the content from the *.nb.ca* domain. This domain holds the provincial education websites, most of which are written in standard French. Additionally, the *.nb.ca* domain is relatively small compared to some other domains (e.g., *.ca*, *.uk*). For the measures that are significant, there is a roughly even split between those that are higher for this corpus vs. the reference corpus.

The Reddit corpus, r/acadie, showed the most significant Chiac measures and the most signifi-

	Acadian French Measures	AcadianDictionary	BootCaT	CC_NB_Domain	r/acadie
Acadian	Acadian French Tokens				
	Acadian French Types				
	Auxiliary *Avoir* Ratio				
	Acadian Conjunctions				
	Acadian Relative Pronouns				
	Questions Containing *ti*				
	English Verb Tokens				
	English 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				
	Third Person Plural *-ont* Verbs				
	-ly Adverb Tokens				
	-ly Adverb Types				
	Instances of *you know*				
	Instances of *right*				
	Instances of *back*				
	Instances of *own*				

Table 7: Comparison of each target corpus to the frWaC.

cant measures overall. The only measure which is higher for the reference corpus is Acadian Relative Pronouns; however, in the analysis in Section 5.4, this measure was not found to be higher for Acadian French than for standard French. As such, this might not be an indication that r/acadie is not Acadian. Despite the strong indications that this corpus is Acadian, it is also the smallest of the Acadian corpora constructed.

The findings of this analysis, along with those for keyword analysis in Section 5.2, indicate that all three corpora — BootCaT, CC_NB_Domain, and r/acadie — hold Acadian French content within them.

6. Conclusion

This work examined how off-the-shelf NLP tools perform on Acadian French text. Our findings showed that they perform worse on Acadian French text than on standard French text. This confirmed the importance of an Acadian French corpus, which could potentially be used for training NLP systems to improve their performance on Acadian French.

We created three corpora using web-as-corpus methodologies. The first was a domain-crawled corpus. We used CommonCrawl's extensive database and scraped New Brunswick (*.nb.ca*) websites. The second was a social media-based corpus from Reddit, where we gathered all posts and comments in the "r/acadie" subreddit. The third was a BootCaT-based corpus, where we used Acadian French words found within the other two corpora to form queries which were sent to a search engine to retrieve webpages containing those terms.

We proposed 22 statistical corpus-based measures based on Acadian French characteristics found in previous research. We showed that these measures indicate that Acadian French text is more Acadian than standard French text. Using these measures, and keyword analysis, we compared our three newly-built corpora to frWaC. Our findings showed that all three corpora hold Acadian French content within them. We found that r/acadie had the highest number of Acadian characteristics while BootCaT was the largest corpus.

This work was partially limited by the lack of current-day NLP tools trained in Acadian French. This is the case since some of our measures depended on correctly tagged words, which we showed, in Section 3.3, the POS tagger worked less accurately on Acadian French text. However, the analysis in Section 5.4 is motivated in part by such limitations to understand their impact on the behaviour of these measures, with the findings in Table 6 suggesting that the measures can be used to identify Acadian French.

A possibility for future work would be to confirm that these corpora would help off-the-shelf NLP tools to perform better on Acadian French. This would help address the aforementioned limitation. Additionally, researchers could use Twitter as another source of social media text because of the availability of information about the geographic location of the origin of tweets. Equivalently, it may be interesting to investigate how

the measures in our research differentiate Acadian French from closely related French varieties (e.g., Quebec French, Fransaskois (Saskatchewan French), Franco-Ontarian French). Depending on the French variety, it may have a harder time distinguishing Acadian French. Alongside this, researchers could create measures to distinguish these closely related French varieties from standard French.

7. Bibliographical References

Anne Abeillé, Lionel Clément, and Alexandra Kinyon. 2000. Building a treebank for French. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*, Athens, Greece. European Language Resources Association (ELRA).

Anne Abeillé, Lionel Clément, and Loïc Liégeois. 2019. Un corpus arboré pour le français: le french treebank [a parsed corpus for french: the french treebank]. *Traitement Automatique des Langues*, 60(2):19–43.

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

Khetam Al Sharou, Zhenhao Li, and Lucia Specia. 2021. Towards a better understanding of noise in natural language processing. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 53–62, Held Online. INCOMA Ltd.

Wedad Altawel. 2021. *La langue comme véhicule identitaire: analyse linguistique du Brayonnaire (Madawaska, Nouveau-Brunswick, Canada)*. Universite de Moncton (Canada).

Laurence Arrighi. 2007. 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*, (57):47–56.

Laurence Arrighi. 2014. Le français parlé en acadie: description et construction d'une «variété». *Minorités linguistiques et société*, (4):100–125.

Patricia Balcom, Louise Beaulieu, Gary R Butler, Wladyslaw Cichocki, and Ruth King. 2008. The linguistic study of acadian french. *Canadian Journal of Linguistics/Revue canadienne de linguistique*, 53(1):1–5.

Timothy Baldwin, Paul Cook, Marco Lui, Andrew MacKinlay, and Li Wang. 2013. How noisy social media text, how diffrnt social media sources? In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 356–364, Nagoya, Japan. Asian Federation of Natural Language Processing.

Marco Baroni and Silvia Bernardini. 2004. BootCaT: Bootstrapping corpora and terms from the web. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

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

Tommy Berger. 2020. 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. Accepted: 2021-01-11T16:32:29Z.

Henri Biahé. 2017. *PARLERS HYBRIDES EN TRADUCTION: L'EXEMPLE DU CHIAC ET DU CAMFRANGLAIS*. Ph.D. thesis.

Andreas Blombach, Natalie Dykes, Philipp Heinrich, Besim Kabashi, and Thomas Proisl. 2020. A corpus of German Reddit exchanges (GeRedE). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6310–6316, Marseille, France. European Language Resources Association.

Albert Branchadell. 2011. Minority languages and translation. *Handbook of translation studies*, 2:97–101.

Peter F Brown, John Cocke, Stephen A Della Pietra, Vincent J Della Pietra, Frederick Jelinek, John Lafferty, Robert L Mercer, and Paul S Roossin. 1990. A statistical approach to machine translation. *Computational linguistics*, 16(2):79–85.

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

Jamie Callan, Mark Hoy, Changkuk Yoo, and Le Zhao. 2009. Clueweb09 data set.

Ian Campbell. 2007. Chi-squared and fisher–irwin tests of two-by-two tables with small sample recommendations. *Statistics in medicine*, 26(19):3661–3675.

Marine Carpuat. 2014. Mixed language and code-switching in the Canadian Hansard. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*, pages 107–115, Doha, Qatar. Association for Computational Linguistics.

Wladyslaw Cichocki, Sid-Ahmed Selouani, and Louise Beaulieu. 2008. The racad speech corpus of new brunswick acadian french: design and applications. *Canadian Acoustics*, 36(4):3–10.

William G Cochran. 1954. Some methods for strengthening the common χ 2 tests. *Biometrics*, 10(4):417–451.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Paul Cook and Laurel J. Brinton. 2017. Building and evaluating web corpora representing national varieties of english. *Language Resources and Evaluation*, 51(3):643–662.

Yves Cormier and Russon Wooldridge. 2000. Dictionnaire du francais acadien. *University of Toronto Quarterly*, 70(1):172.

Mark Davies and Robert Fuchs. 2015. Expanding horizons in the study of world englishes with the 1.9 billion word global web-based english corpus (glowbe). *English World-Wide*, 36(1):1–28.

Guy De Pauw. 2006. Developing Linguistic Corpora—A Guide to Good Practice. *Literary and Linguistic Computing*, 22(1):101–102.

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

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

Andreas Eisele and Yu Chen. 2010. MultiUN: A multilingual corpus from united nation documents. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

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

W. N. Francis and H. Kucera. 1979. Brown corpus manual. Technical report, Department of Linguistics, Brown University, Providence, Rhode Island, US.

Karine Gauvin. 2004. Bdlp·acadie. Last accessed 24 février 2023.

Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. 2012. Building large monolingual dictionaries at the Leipzig corpora collection: From 100 to 200 languages. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 759–765, Istanbul, Turkey. European Language Resources Association (ELRA).

Barry M Gough. 2010. *Historical dictionary of Canada*. Scarecrow Press.

Sidney Greenbaum and Gerald Nelson. 1996. The international corpus of english (ice) project. *World Englishes*, 15(1):3–15.

Naomi ES Griffiths. 1992. *Contexts of Acadian history, 1686-1784*. McGill-Queen's Press-MQUP.

Emiliano Raul Guevara. 2010. NoWaC: a large web-based corpus for Norwegian. In *Proceedings of the NAACL HLT 2010 Sixth Web as Corpus Workshop*, pages 1–7, NAACL-HLT, Los Angeles. Association for Computational Linguistics.

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

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

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

Adam Kilgarriff. 2009. Simple maths for keywords. In *Proc. Corpus Linguistics*, volume 6.

Adam Kilgarriff, Siva Reddy, Jan Pomikálek, and Avinesh PVS. 2010. A corpus factory for many languages. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

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

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand.

H. Kucera and W. N. Francis. 1967. *Computational analysis of present-day American English*. Brown University Press, Providence, RI.

Geoffrey Leech et al. 1992. 100 million words of english: the british national corpus (bnc). *Language research*, 28(1):1–13.

Vinci Liu and James R. Curran. 2006. Web text corpus for natural language processing. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 233–240, Trento, Italy. Association for Computational Linguistics.

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

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, Baltimore, Maryland. Association for Computational Linguistics.

France Martineau. 2011. 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.

France Martineau. 2014. L'acadie et le québec: convergences et divergences. *Minorités linguistiques et société / Linguistic Minorities and Society*, (4):16–41.

France Martineau and Marie-Claude Séguin. 2016. Le corpus fran: réseaux et maillages en amérique française. *Corpus*, (15).

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

Aaron Mueller, Garrett Nicolai, Panayiota Petrou-Zeniou, Natalia Talmina, and Tal Linzen. 2020. Cross-linguistic syntactic evaluation of word prediction models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5523–5539, Online. Association for Computational Linguistics.

Brian Murphy and Egon W. Stemle. 2011. PaddyWaC: A minimally-supervised web-corpus of hiberno-English. In *Proceedings of the First Workshop on Algorithms and Resources for Modelling of Dialects and Language Varieties*, pages 22–29, Edinburgh, Scotland. Association for Computational Linguistics.

Marie-Ève Perrot. 1995a. *Aspects fondamentaux du métissage français/anglais dans le chiac de Moncton (Nouveau-Brunswick, Canada)*. Ph.D. thesis, Paris 3.

Marie-Ève Perrot. 1995b. Tu worries about ça, toi? métissage et restructurations dans le chiac de moncton. *Linx*, 33(2):79–85.

Marie-Ève Perrot. 2001. Bilinguisme en situation minoritaire et contact de langues: l'exemple du chiac. *Faits de langues (Evry)*, (18):129–137.

Marie-Ève. Perrot. 2014. Le trajet linguistique des emprunts dans le chiac de moncton: quelques observations. *Minorités linguistiques et société / Linguistic Minorities and Society*, (4):200–218.

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

Pascal Poirier and Pierre Marie Gérin. 1993. *Le glossaire acadien*. Moncton, N.-B.: Éditions d'Acadie.

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

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

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Randolph Quirk. 1990. Language varieties and standard language. *English today*, 6(1):3–10.

Ella Rabinovich, Yulia Tsvetkov, and Shuly Wintner. 2018. Native language cognate effects on second language lexical choice. *Transactions of the Association for Computational Linguistics*, 6:329–342.

Alan Ritter, Colin Cherry, and Bill Dolan. 2010. Unsupervised modeling of Twitter conversations. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 172–180, Los Angeles, California. Association for Computational Linguistics.

Roland Schäfer and Felix Bildhauer. 2013. Web corpus construction. *Synthesis Lectures on Human Language Technologies*, 6(4):1–145.

James Calvert Scott. 2004. American and british business-related spelling differences. *Business Communication Quarterly*, 67(2):153–167.

Charlene Soucy-Godby and Gert Michaud. 2014. Brayonnaire: Petit dictionnaire brayon / français / anglais.

Charlene Soucy-Godby and Gert Michaud. 2016. Brayonnaire – partie 2: Petit dictionnaire brayon / français / anglais.

Johansson. Stig, Geoffrey N. Leech, and Helen Goodluck. 1978. *Manual of information to accompany the Lancaster-Oslo : Bergen Corpus of British English, for use with digital computers*. Department of English, University of Oslo.

Vít Suchomel and Jan Pomikálek. 2012. Efficient web crawling for large text corpora. In *Proceedings of the seventh Web as Corpus Workshop (WAC7)*, page 39–43.

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

Spencer Trerice. 2016. *Entre fierté et mépris: le rapport ambivalent à l'égard du chiac dans 'Pour sûr' de France Daigle*. Ph.D. thesis.

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

Graham J. G. Upton. 1982. A comparison of alternative tests for the 2 × 2 comparative trial. *Journal of the Royal Statistical Society. Series A (General)*, 145(1):86–105.

Graham JG Upton. 1992. Fisher's exact test. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 155(3):395–402.

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

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