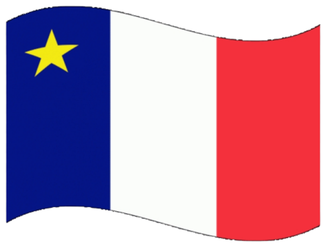

WaCadie: A Web Corpus of Acadian French

Jérémy Robichaud and Paul Cook
The University of New Brunswick
{jrobich9,paul.cook}@unb.ca

UNB EST. 1785 UNIVERSITY OF NEW BRUNSWICK

1. Research Questions

- Can we create an Acadian French corpus using previously proposed Web-as-Corpus methodologies?
- If so, is there one methodology that surpasses others in quality and quantity?

2. Acadians

- Acadians are a linguistic minority in the Atlantic region of Canada, with a large population in New Brunswick.
- Some sub-regions of New Brunswick have different variants of Acadian French.
  - **Acadian French**: All regions
  - **Brayon**: North-West regions
  - **Chiac**: South-East regions

3. Known Properties

- Previous Linguistic research has identified properties of:
- **Acadian French**:
  - Acadian-only words (*Harder, Picasse, Frolic*)
  - English variants of French verbs (*parker, winker*)
- **Brayon**:
  - Brayon-only words (*Bagosse, Jnou, arrâs*)
  - Brayon expressions (*Chu brûlé bin tight!*)
- **Chiac**:
  - English words (*Je mange des fries*)
  - Non-standard spelling of 3rd-person plural French verbs (*ils parlont*)
- From these, we created 22 computational measures to identify the Acadian French properties of any corpus.

4. Web-as-Corpus

- We can leverage the web to create corpora of languages using multiple approaches:
  - **BootCaT**: Corpus from websites found on a search engine (Google) while searching for Acadian French words.
  - **Domain Crawling**: Corpus of .nb.ca domain websites using snapshots from the web.
  - **Social Media**: Corpus from the posts and comments on the r/acadia subreddit.

5. Pipeline

- We post-process the data to standardize and clean it.
  - **File Encoding**: Encode with UTF-8.
  - **Justext**: Retrieves texts from HTML.
  - **Accent Normalization**: Remove all accents.
  - **Junk Filter**: Remove non-French Sentences.
  - **Exact Dedup**: Removing duplicate documents using hashes.
  - **PyOnion**: Near Deduplication removal.
  - **Stanza**: Handles tokenizing, part-of-speech tagging, lemmatization, and dependency parsing.

6. Conclusion

- **Results**:
  - All corpora created have many more Acadian properties than a standard French web corpus.
- **Future Work**:
  - Update the tests to better reflect the data found.
  - Create a Twitter corpus.
  - Conclude if the created corpora are Acadian French.