Digitizing Labrador Languages

The problem

- The Indigenous languages of Labrador are under threat
- Inuttut, spoken by the Labrador Inuit people, is highly endangered, with only a few hundred speakers left
- Innu-aimun, spoken by the Innu people (unrelated), is in better shape, but fewer and fewer kids are learning the language every year
- There is an urgent need for language teachers and teaching materials

The problem

- Linguists have been documenting these languages for decades
- There have also been numerous newsletters, storybooks, etc., published over the years in these languages
- A huge amount of material exists in MUN archives
- Making it accessible as an online searchable database would make it useful to language teachers and curriculum designers in Labrador Indigenous communities
- Large quantities have not been digitized
- What has been digitized mostly exists as image files
- Some has been laboriously transcribed by hand

Pîtshâu nete tshematet. Ispânâu ne mi'tshuâp. Kue pî'tsheiân. Shâkuâshu nte pemu'teiân. Kue etu'teiân nitapunit. Takuan nimashinaikan mâk nimashinaikanâskua.

Mekuat aiamu katipenitak, uitamu eshpish minuenitak minuat uiapamat katshishkutamuakanishiniti.

Tshekimaun ntapestan ua tshipitiman nuiashem kie eukun iapesteian emitsheshian.

(1) Ntshen: Ut[illegible]maut nte Ottawa kie kamassit kie [illegible]estukuiat, esk apu sentak ne auen [illegible]shipa [illegible]pentamut nenu kamanishan[illegible]shi Ausseuna.

Developing OCR tools

- Automating the process would make it much faster
- Optical Character Recognition
- Numerous off-the-shelf OCR applications exist: Adobe Acrobat, ABBYY FineReader, etc.
- The higher-end apps, like FineReader, support adding additional languages
- However, this is often less than satisfactory
  - Little support for glyph variation (a **ɑ** **ɑ** *a*)
  - Inadequate support for non-standard characters
  - No support for probability of collocations (*qu* vs. *qn*, etc.)

— Leshup ulaken —

Leshup ulaken apestakenu estaken leshup, kalasteshet kie mak kapeleshkueu apui.

Lesf-i p ula ken

Lesk up ulaken ctpestakenu estaken iesh м p ? kalasteshet kie mak kapele^h kueu dpui .

Developing OCR tools

- Solution: develop new OCR suites tailored to the individual languages and to the materials on hand
- OCR Tesseract is a tool for doing this
  - allows building a large set of variations for each glyph by scanning
  - can analyze text statistically for collocation probabilities
  - implemented in Python

The project

- Develop a working OCR for each language (Innu-aimun & Inuttut)
- Working with existing textual materials to build libraries of glyphs, variants, and collocations (*tsh, k^{u}, iK,* etc.)
- Testing it on other texts
- Some coding, in Python, might be involved
- Dr. Wareham has suggested this work could be a term project for this class

Complications

- I'd like to get an accuracy of at least 95% to minimize manual clean-up
- This is challenging because
  - Some texts are dual-language: Innu-French, Innu-English, Inuttut-English, Inuttut-German
  - Some contain both handwritten text and typed/printed text
  - Many use obsolete spelling systems
  - Many contain spelling or transcription errors
- Therefore
  - The larger the amount of textual input, the better (eventually, statistical trends should overwhelm noise)
  - Ditto for glyph variations: fonts/typefaces, handwriting samples

The follow-up

- Success in this project will be followed by another (probably next term):
- developing OCR for Iroquoian and Dene languages
  - more complicated character sets
- ultimately, moving on to non-Roman writing systems (chiefly Aboriginal Syllabics)
- Those who have developed the Inuttut and Innu-aimun projects may be hired as research assistants for subsequent projects