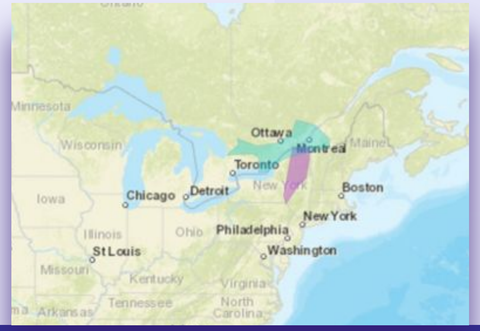
Creating a Corpus of Kanien’kéha (Mohawk)

Lessons learnt while creating a corpus for a (mostly) oral language

Akwiratékha Martin*, Deirdre Demson*♦, Thanyehténhas Nathan Brinklow✤ and Anna Kazantseva*

Kanien’kéha

- Iroquoian language spoken in parts of North Eastern US, and Quebec and Ontario in Canada
- Polysynthetic language with rich morphology. Verbal morphology is especially elaborate (16 prefixes consisting of non-modal prepronominals, tense and pronominals and reflexives. 16 suffixes consisting of aspects, and root suffixes like inchoative, causative, benefactive; affixes have up to 9 forms).
- Speakers: 500-600 first language (L1) speakers, 75 second language (L2) speakers (Advanced-Mid proficiency), about 25 L1 speakers raised by L2 parents, unattested number of L2 speakers with Novice to Intermediate proficiency over a thousand.
- Example:
  Thiaonsakoniatatshennaranì:ron.
  Th-i-a-onsa-koni-atat-shenn-a-r-ani-hr-on
  Contras-transloc-opt-rep-1st.2nd.sing-reflex-noun.name- link-verb.contain-ben-purp-stat
  '*I should have just gone and signed your name*'.



- 4 main dialects (spoken in 7 communities).
- Few accessible and searchable written resources, no corpora, no complete grammar.

Motivation

- First available corpus.
- Studying usage patterns.
- Usage for creating language technology.

Objectives

- Unify orthography.
- Correcting morphological accuracy and tone.
- Transliteration: making parallel versions available in different dialects of Mohawk.

Corpus

- A small corpus: 100,014 tokens.
- 68 documents : 51 Bible chapters, 13 movie scripts, 4 others.
- Type-token ratio (ratio of all words to unique words) is 3.75.
- Average word length is 7.15 characters.
- Average sentence length is 11.57 words

Methodology

Unification of Orthography

- Original files are in the Eastern (Kahnawà:ke and Kanehsatà:ke) dialect written over the last 30 years.
- Spelling mostly follows the Mohawk Language Standardization Project
- However, orthography is inconsistent even within dialects.
- All documents are manually proofread by an advanced level L2 speaker correcting inconsistencies.
- Original files/versions preserved.
- Frequent errors: morphological mistakes, nouns and aspect suffixes lacking glottal stops. Tone, falling tone on lengthened vowels followed by a glottal stop.

Transliteration

- Corpus was automatically translated into the Eastern dialect.
- We could not automatically translate into the older Western dialect: inaccuracies would outweigh the benefits

Contributions

- First corpus of Kanien’kéha for linguistic and computational research.
- Small but high quality corpus (manually proofread).
- Available in two Eastern dialects.
- Preliminary analysis completed.

Future Work

- Continued proofreading.
- Expansion of corpus material from community archives and Canadian and USA archives.
- Evaluation of existing language technology for Kanien’kéha.

Positionality

The choice of content, the linguistic guidance and expertise involved in this project are provided by Mohawk community members, Kanien’kehá:ka. Their work is based on decades of language revitalization efforts. The funding, infrastructure and technical expertise are by the National Research Council Canada.

CONTACT:
Akwiratekha.Martin@cnrc-nrc.gc.ca
Anna.Kazantseva@nrc-cnrc.gc.ca
nathan.brinklow@queensu.ca
deirdredemson@gmail.com

*National Research Council Canada, ✤Queen’s University, ♦SUNY at Buffalo