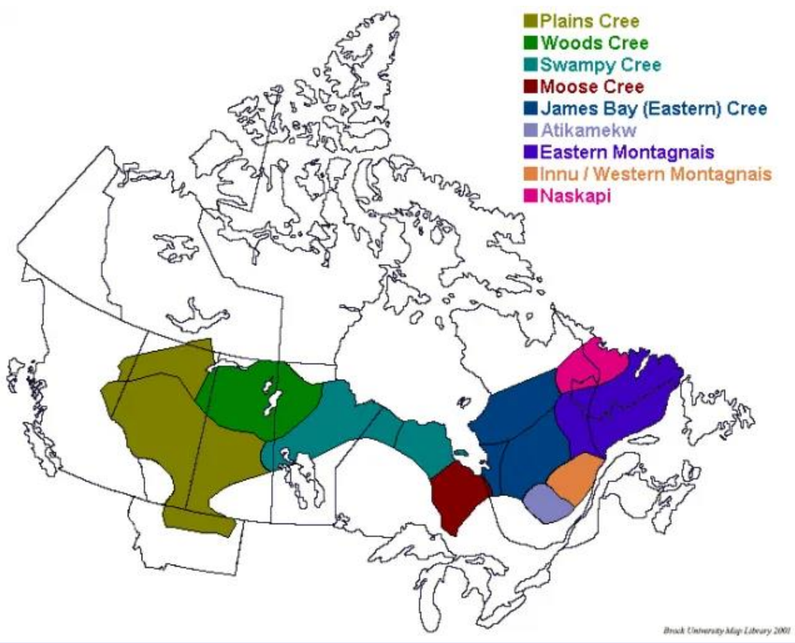
Building a Language Model of nehiyawewin (Cree, Y-dialect)

Delaney Lothian, Daniela Teodorescu, Gokce Akcayir, Denilson Barbosa, Carrie Demmans Epp

Department of Computing Science, University of Alberta

What is nehiyawewin?
ᓀᐦᐃᔭᐍᐏᐣ

- Cree is a polysynthetic language that is Indigenous to the plains of Canada.
- Cree is a low resource language.



Low-Resource Languages

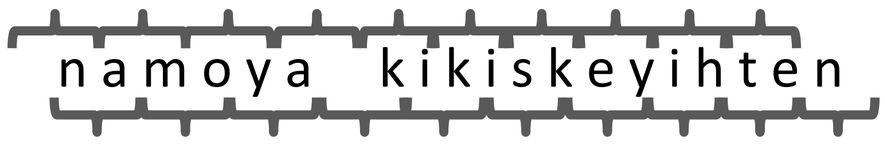
- Many companies ignore languages like Cree because there are few resources, which means they cannot create tools to support these languages.
- For learners of Indigenous languages, this lack of effective and engaging resources is a common barrier to learning.
- Technology can help us overcome this barrier, provided language structure and cultural elements are used to support ancestral language learning.
- To provide this support, we need automated approaches to describing the language

We Need A Cree Language Model

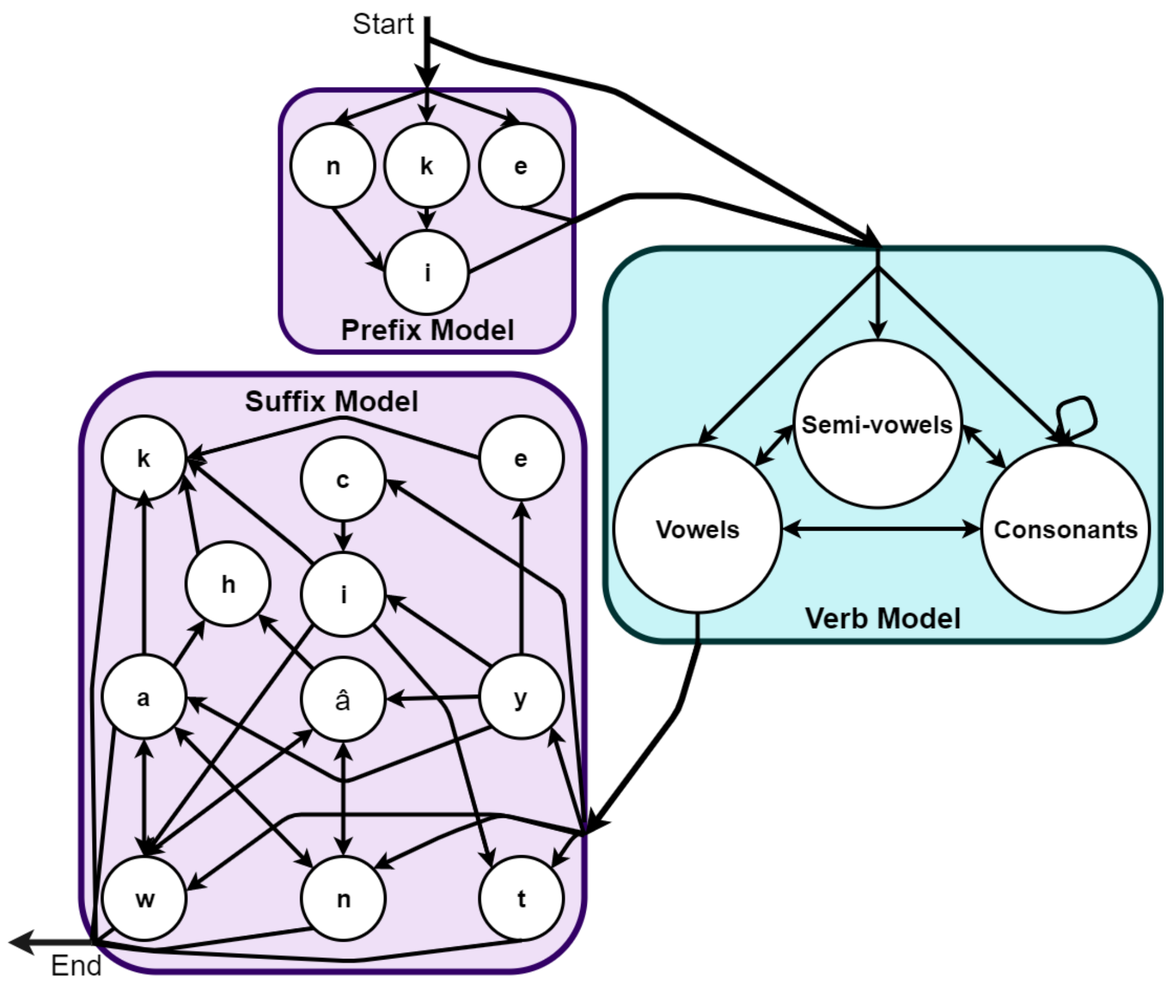
- To create quality resources to aid in the revitalization of Cree and other Indigenous languages, we need a representation of the structure of the language.
- This representation must be understandable and usable by computers.
- These representations are called language models.
- Language models are essential to supporting the creation of other language technologies (e.g., speech recognition).

Character Level N-Gram Language Model

Bigrams



Language Model (Animate Intransitive Verbs)



Verb Conjugation Decomposition

Having built the language model, the program will first pass through the given Cree text forwards to find the prefix

ki mîcisonâwâw →

And then pass through it backwards to find the suffix.

kimîciso nâwâw ←

Having found the affixes, the program assumes the remaining text in the middle is the verb.

The next step in this task will be the morphological analysis of nouns. Consider the word for horse, **mistatim.**

mist/atim
Big Dog

This Language Model Will Support the Development of Language Technologies

- The language model will support the creation of tools that will give learners of Indigenous languages opportunities to practice
  - listening, through text-to-speech technologies, and
  - speaking, while receiving feedback from speech-recognition tools.
- Speech recognition will also help address the transcription bottleneck communities are currently experiencing.
- Language-learning systems will be able to offer feedback to support learning.

NSERC CRSNG

UNIVERSITY OF ALBERTA