UNIVERSITY OF ALBERTA

Word Class Frequencies According to Corpora

Katherine Schmirler and Atticus G. Harrigan

March 5, 2016

Outline

- Introduction
  - Who are we?
  - What do we do?
- Background
  - Finite State Machines
  - Purpose of Tools
- Verbal Counts
- Nominal Counts
- Further Development
- Conclusion
- References

Who are we?

- Alberta Language Technology Laboratory (ALT Lab)
- International group of collaborators from within Canada (FNUC) and beyond (Giellatekno at UiT)
- Team of theoretical, corpus, and computational linguists

What do we do?

- Focus on the creation of linguistic tools, especially for understudied languages
- Project working on Algonquian, Siouan, Athabaskan, and Haida languages
- Tools take the form of smart online dictionaries, speech recognition, speech synthesis, corpus construction

Finite State Machines

- We make use of finite state tools to perform morphological analysis and generation
- A finite state machine is one where an underlying form is transformed into a surface form. These machines are reminiscent of basic phonological rewrite rules; given a rule where X => Y / W __ Z, we could create the following machine:

Surface	W	Y	Z
	|	|	|
Underlying	W	X	Z
Result	wyz		

- We can expand our machines to transduce grammatical tags into morphemes (e.g. Plains Cree):

Surface	a	t	i	m	ε	ε	ε	ε	ε	wak
	|	|	|	|	|	|	|	|	|	|
Underlying	a	t	i	m	+	N	+	AN	+	Pl
Result	atimwak									

Finite State Machines

- **We may add a layer of phonological transformation**
- **For example, a simple Plains Cree diminutive where a noun is suffixed with -{(s)is} while and /t/ in the word are affricated into [t͡s]**
- **We set this as a rule to occur with our diminutive tag (+Der/Dim):**

Surface	a	c	i	m	ε	ε	ε	ε	ε	osis	ε	ε	ε	ε	ε	ε
	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|
Underlying	a	t	i	m	+	N	+	AN	+	Der/Dim	+	N	+	AN	+	Sg
Result	acimosis															

Finite State Machines

- **Using combinations of these machines, we have created a morphological analyser:**

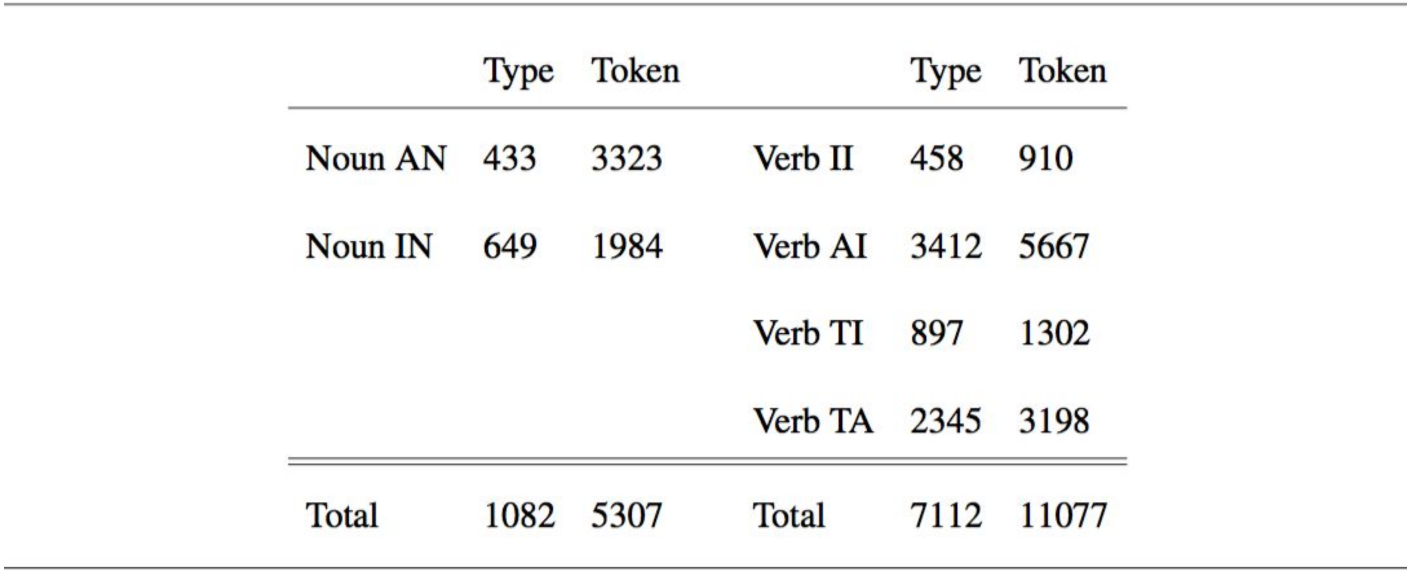
  Input String: ê-nimihitoyân

  Analysis: PV/e+nîmihitow+V+AI+Cnj+Prs+1Sg
- **We can use this tool to automatically analyse large bodies of text and/or corpora**
- **Using a corpus of conversational Plains Cree provided to us by Dr. Wolfart, we performed such an analysis**

Counts

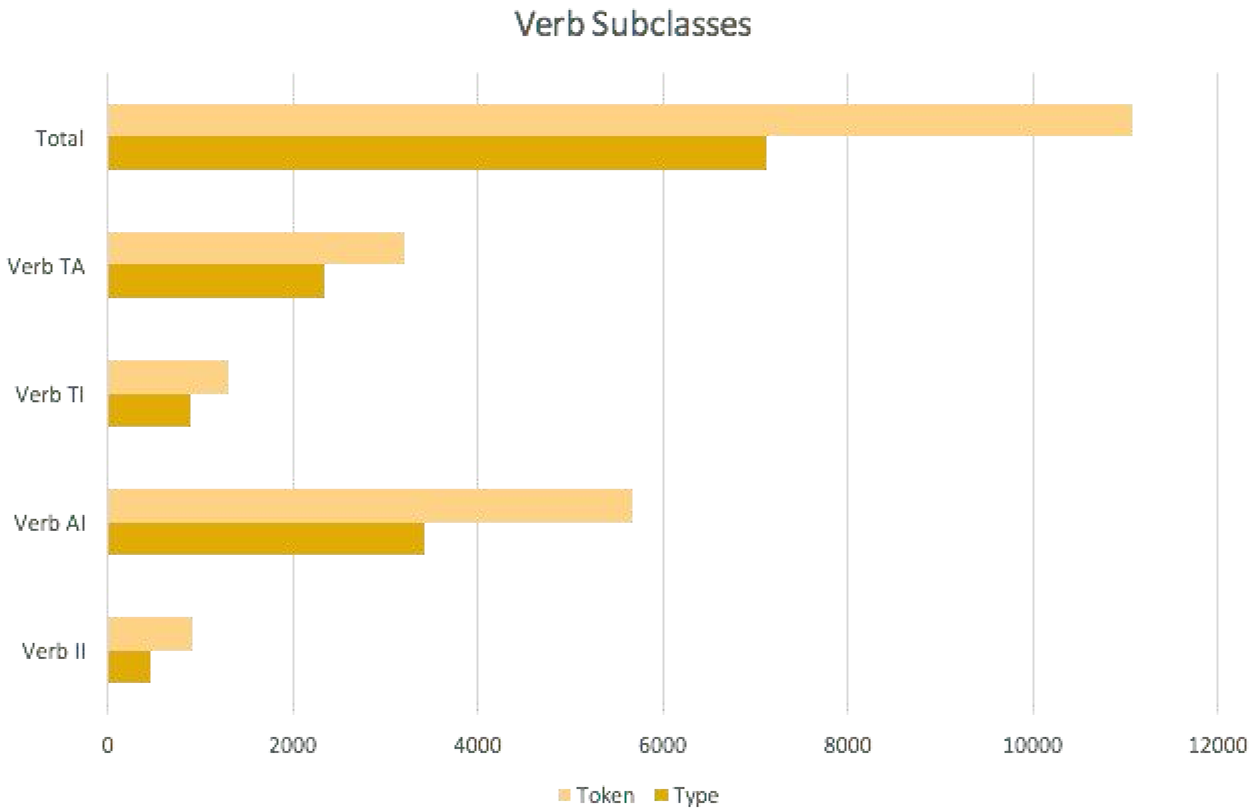
- The overall corpus contains 18,646 types representing 125,368 tokens
- 8606 types (32,399 tokens) were unanalysed, leaving 10,040 types and 92, 969 tokens analysed
- Punctuation accounted for 40,560 tokens, leaving 52,409 tokens as linguistic units

Counts

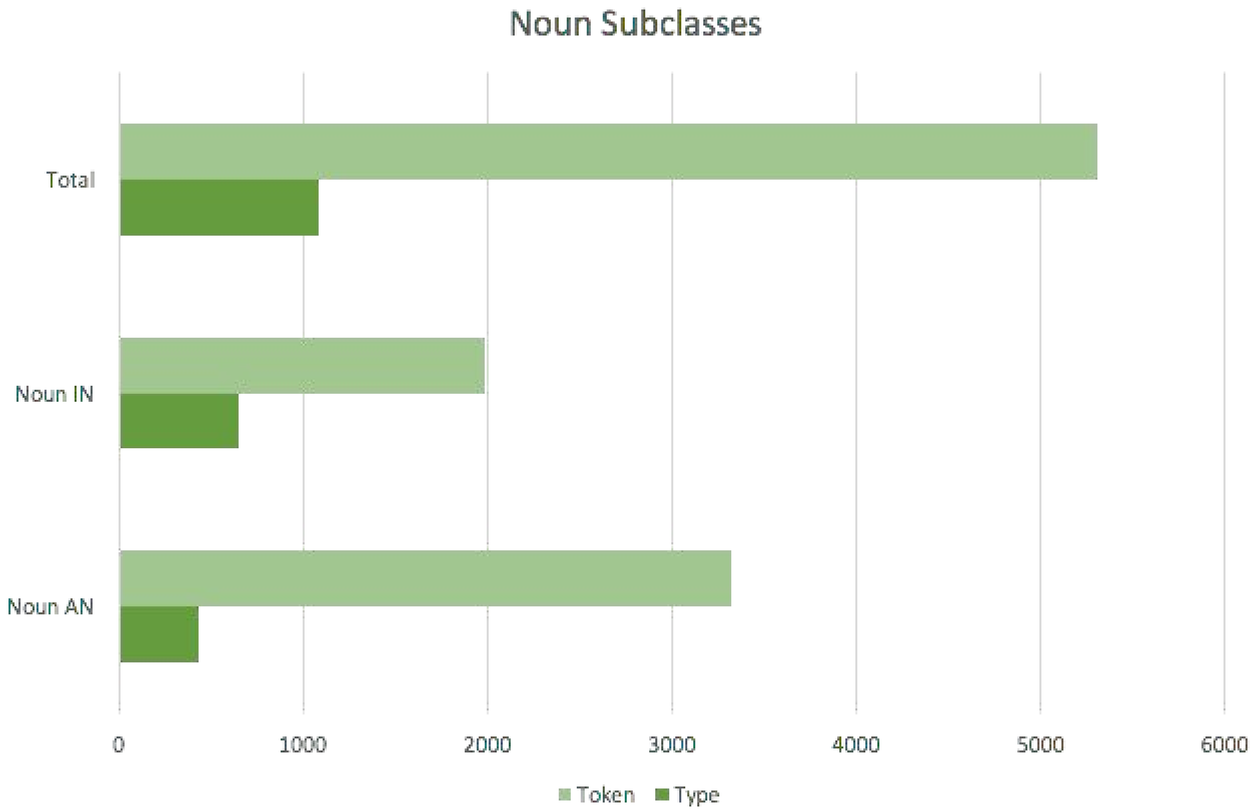


	Type	Token		Type	Token
Noun AN	433	3323	Verb II	458	910
Noun IN	649	1984	Verb AI	3412	5667
			Verb TI	897	1302
			Verb TA	2345	3198
Total	1082	5307	Total	7112	11077

Verb Counts


Verb Subclasses
Total
Verb TA
Verb TI
Verb AI
Verb II
0
2000
4000
6000
8000
10000
12000
Token
Type

Noun Counts


Noun Subclasses
Total
Noun IN
Noun AN
0
1000
2000
3000
4000
5000
6000
🟩 Token
🟩 Type

Caveats

- These counts are for unambiguous analyses
    - If our system could break down the word into morphemes in more than one way, it was not counted here (e.g. the obviative singular and plural forms)
    - This is because many of our ambiguous analyses are due to generous descriptions
    - Some ambiguity is valid, but is not covered here
- 8606 types (32,399 tokens) were unanalysed, leaving 10,040 types and 92, 969 tokens analysed
    - ongoing hand-verification
- Punctuation accounted for 40,560 tokens, leaving 52,409 tokens as linguistic units

Adaptation to other dialects

- **Basic infrastructure already there, just needs to be adapted**
- **Assumptions:**
    - **Western Cree dialects differ in terms of a few key sounds alongside lexical differences**
    - **Standard Roman Orthography is used for all**
    - **E.g. Woods Cree:**
        - **<th> recognised as Plains Cree <y>**
        - **<ī> can be recognised Plains Cree <ē>**

Initial attempts at recognising Woods Cree

- In collaboration with Miikka Silfverberg
- A simple Woods Cree text (a story from Solomon Ratt, transcribed in the SRO by me)
    - 324 words
- Without any spelling rules: 60% recognised
    - i.e., 60% of the words and spellings are identical to Plains Cree
- Spelling change rules:
    - Recognise <th> as <y>, recognise <ī> as <ē>
    - Loosened rules for vowel length before <y, w, h> due to errors in my transcription
- With these rules: 72% recognised

What does this leave unrecognised?

- Systematic lexical differences
  - Woods *ikwa* vs. Plains *ēkwa* and compounds
    - *ikota*, *ikotī*, *ikospī*, *ikwāni*, etc.
- Plains Cree items that haven’t made it into our lexicon
  - *wīwa* ‘his wife’
  - *piko* variant *poko*
- Transcription errors
  - Primarily *Vh* instead of $\bar{V}$

Further considerations

- This only allows for *recognition* of Woods Cree by loosening spelling rules
  - These rules generate dozens of incorrect forms
- Steps to take:
  - Include both types of <y> (*y* and *ý* in Wolvengrey 2001) in the Plains Cree model
    - <y2> becomes <th> in Woods Cree to generate and <th> becomes <y2> for recognition
  - But <ī> and <ē> differences work differently
    - Every Plains <ē> becomes <ī> but not every Wood <ī> becomes <ē>
  - Increase the lexical items available
  - Standardising orthography
- Thanks to Solomon Ratt for giving us access to more texts to continue development

Derivational morphology

- In collaboration with Antti Arppe
- In the very beginning stages of development
  - Thanks to Arok Wolvengrey for his work on derivations for his dictionary entries, and for supplying them to us
- Challenges:
  - Morphophonological rules
    - w+i > o / C __
    - Vy+i/Vw+i > V̄
    - t > c / __ i (but not every *i*) – historical considerations
  - Overgeneration of possible analyses
    - Weighted analyser chooses the simplest so far, but this is not always going to be correct

Conclusions

- **Using a basic corpus of only ~100,000 words, we can observe a few interesting patterns for Plains Cree word classes:**
  - **Verbs are more common than nouns, both in types and tokens.**
  - **The most frequent verb class is VAI, followed by VTA, then VTI, and finally VII.**
  - **Animate nouns have more token than inanimate nouns, though fewer types.**
    - **Combined with the above, this makes sense: It seems our corpus focused on the animate.**
  - **Through further computational development and guided by grammatical descriptions, we plan to expand our system to recognize more of the language, especially derivational processes**
  - **We further plan to expand our system to other dialects of Cree**

References

Ahenakew, Freda, ed. 1987. *wāskahikaniwiyiniw-ācimowina / Stories of the House People, Told by Peter Vandall and Joe Douquette.* Winnipeg: University of Manitoba Press.

Ahenakew, Freda and H.C. Wolfart, eds. 1997. *kwayask ē-kī-pē-kiskinowāpatihicik / Their Example Showed Me the Way: A Cree Woman's Life Shaped by Two Cultures*. Told by Emma Minde. Edmonton: University of Alberta Press.

_____. 1998. *kōhkominawak otācimowiniwāwa / Our Grandmothers' Lives as Told in Their Own Words*. Regina: Canadian Plains Research Center.

Okimāsis, Jean. 2004. *Cree: Language of the Plains / nēhiyawēwin: paskwāwi-pīkiskwēwin*. Regina: Canadian Plains Research Center.

Wolfart, H. Christoph. 1973. *Plains Cree: A Grammatical Study*. Transactions of the American Philosophical Society New Series, vol. 63 (5). Philadelphia: The American Philosophical Society.

_____. 1996. "Sketch of Cree, an Algonquian Language." *Handbook of North American Indians* 17:390-439.

Wolfart, H. C., and Freda Ahenakew, eds. 1993. *kinēhiyawiwininaw nēhiyawēwin / The Cree Language is Our Identity: The La Ronge Lectures of Sarah Whitecalf*. Winnipeg: University of Manitoba Press.

_____. 1998. *ana kā-pimwēwēhahk okakēskihkēmowina / The Counselling Speeches of Jim Kā-Nīpitēhtēw*. Winnipeg: University of Manitoba Press.

_____. 2000. *âh-âyîtaw isi ê-kî-kiskêyihtahkik maskihkiy / They Knew Both Sides of Medicine: Cree Tales of Curing and Cursing Told by Alice Ahenakew*. Winnipeg: University of Manitoba Press.

_____. 2010. *piko kîkway ê-nakacihtât: kêkêk otâcimowina ê-nêhiawastêki*. Winnipeg: Algonquian and Iroquoian Linguistics.

Wolvengrey, Arok. 2001. *nēhiyawēwin: itwēwina / Cree: Words*. Vols. 1 & 2. Regina: Canadian Plains Research Center.

Thank you!

Questions?

schmirle@ualberta.ca
galvin@ualberta.ca
http://altlab.artsrn.ualberta.ca/