Community-based Corpus Building: Three Case Studies

Sally Rice (Linguistics)
Dorothy Thunder (Native Studies)
University of Alberta
Edmonton, Alberta, Canada

ICLDC-5  University of Hawai‘i at Mānoa  3 March 2017
ʻulu analogy

multi-purpose, robust, durable, nutritionally complete

http://globalmana.org/portfolio/breadfruit/
‘ulu analogy

multi-purpose, robust, durable, nutritionally complete

http://globalmana.org/portfolio/breadfruit/
The Whys & Hows of Building & Using Indigenous Language Corpora

- Why 1. Because we can.
- Why 2. Because corpora make for better documentation.
- Why 3. Because corpora make precious language data more useful and accessible for different purposes.

- How 1.
- How 2.
- How 3.

- Case 1.
- Case 2.
- Case 3.
The Whys & Hows of Building & Using Indigenous Language Corpora

- **Why 1.** Because we can.
- **Why 2.** Because corpora make for better documentation.
- **Why 3.** Because corpora make precious language data more useful and accessible for different purposes.

- **How 1.** Record. Transcribe. Check. Digitize. (And keep good metadata!)
- **How 2.** Upload text files into an off-the-shelf concordancer.
- **How 3.** Create words lists, exemplify dictionaries, discover recurrent collocations and patterns of usage.

- **Case 1.**
- **Case 2.**
- **Case 3.**
The Whys & Hows of Building & Using Indigenous Language Corpora

- Why 1. Because we can.
- Why 2. Because corpora make for better documentation.
- Why 3. Because corpora make precious language data more useful and accessible for different purposes.

- How 1. Record. Transcribe. Check. Digitize. (and keep good metadata!)
- How 2. Upload text files into an off-the-shelf concordancer.
- How 3. Create words lists, exemplify dictionaries, discover recurrent collocations and patterns of usage.

- Case 1. Stoney (Alexis Nakoda Sioux Nation); SIOUAN
- Case 2. Dene Sųłiné (Cold Lake First Nation); ATHAPASKAN
- Case 3. nâhiyawêwin (Plains Cree); ALGONQUIAN
Algonquian, Athapaskan, & Siouan Languages
Algonquian, Athapaskan, & Siouan Languages

- all polysynthetic (verb is propositional and has multiple affixes doing lots of inflectional and derivational work)

- verb stem is word-medial or word-final

- much allomorphy & suppletion

- few parts of speech

- more linguistic attention paid to morphology than connected discourse

standard corpus mark-up such as

- TAGGING
- PARSING
- LEMMATIZING

is neither easy-to-do nor necessary for modern Unicode-savvy concordancers
The Whys

- Why 1. We can.
- Why 2. We should.
- Why 3. We have much to gain.

A corpus is a machine-readable collection of texts.

Like an archive, a corpus can be a manifestation of good language documentation.

If the purpose of LD is to provide “a comprehensive record of the linguistic practices characteristic of a speech community” (Himmelmann 1998: 166), then a corpus is an excellent means of achieving this in ways readily accessible to speakers.

Corpora are built from samples of language that represent naturally occurring, connected discourse—not word lists or field notes as often go into archives.

The language data in corpora are natural, expandable, multiply accessible, & reusable.
## Basic characteristics of different research instruments for language documentation

Vinogradov 2016: 136 (Table 3)  
Gries & Berez 2015

<table>
<thead>
<tr>
<th>feature</th>
<th>major corpora</th>
<th>documentary corpora</th>
<th>language archives</th>
<th>printed text collection</th>
</tr>
</thead>
<tbody>
<tr>
<td>selectivity of material</td>
<td>+</td>
<td>+</td>
<td>−</td>
<td>+/−</td>
</tr>
<tr>
<td>machine-readable format</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>−</td>
</tr>
<tr>
<td>volume</td>
<td>big/huge</td>
<td>small</td>
<td>big/small</td>
<td>very small</td>
</tr>
<tr>
<td>annotation</td>
<td>+</td>
<td>+/−</td>
<td>−/+</td>
<td>+/−</td>
</tr>
<tr>
<td>searchability</td>
<td>+</td>
<td>+</td>
<td>−/+</td>
<td>−</td>
</tr>
<tr>
<td>balanced subcorpora</td>
<td>+</td>
<td>−/+</td>
<td>−/+</td>
<td>−</td>
</tr>
<tr>
<td>quantitative analysis</td>
<td>+</td>
<td>−</td>
<td>−</td>
<td>−</td>
</tr>
</tbody>
</table>
The Hows

- **How 1.** Create UTF-8 encoded text files (with metadata!)

- **How 2.** Upload them into a concordancer (e.g. AntConc)

- **How 3.** Make wordlists
  - Search for keywords-in-context (KWIC)
  - Discover recurrent fixed expressions (n-grams)
  - Find collocates of an item (be it morpheme, word, or expression)
- See the connection between language use and genre/register
- Build exemplified dictionaries & grammars
Three Case Studies

- Stoney: data collection
- Dene Sųłiné: text rendering & upload
- nêhiyawêwin: corpus queries
Collecting Data for the Corpus

- what kind of corpus?
  snapshot, monitor, opportunistic

- what kind of sources?
  by demographic, by domain

- what types of genres?
  S written stories (edited)
  N traditional stories, personal narratives (planned)
  C conversation (unplanned)

* includes strategic file-naming & maintenance of separate metadata file
Collecting Data for the Corpus

- Collect with a camera, smartphone/tablet, or digital recorder.

**Record**
- Use Audacity/SayMore to save recording as mono .wav or .aiff file.

**Transcribe**
- Transcribe recording using Elan/VoiceWalker.
- Export transcription as .txt file.
Collecting Data for the Corpus

How 1

Yusbecakebiti da yecash Tammy daca ugucu?

Yusbecakebiti da yecash ozorobi sibihnu gici ugucud.
Data File & Metadata Management

HOW 2

Sekwi hesli hu ịa ededoresni sa duwe nį.
Duhu t’ahì dechêngohanai kùwè ho ?ą eyer
nùneshik’odene ba yeh hudelgaì chok ho?ą nį.
Eku Jethk’e desaze nask’edhe tsąba nalye nį ?ānį.
Yų nānè nįbali aleza nadathe ?ai tth’i dene aleza dzeredił hutow
elgha kùta delth’hì kun bąlghe.
Dighi naghelt’e tli tth’i hedonelt’u nîbâli kùwè hoga dathetą.
Lîchok hoyini t’adore ?a nį ekuwu.
Taghe yį theni nagheț’elį dathela.
Ițlaghe tsąba k’odhere bebethchêne hu, țlaghe nedîghanai,
eku țlaghe belaye kun debadhì dathe ?ą
eyi bețihe delk’osi bebethchêne nį.
Eku tth’i lîchok ts’į beyatt’h’aghe elk’ech’a
adadaghelyi darîth’h’ak.

Eku seta nį tthetheyi hîlî?ą eyer tsąba k’odhere chu be?ihe
delk’osi xeł theda hot’a seghâ yati thet ą nį.
Eyer dene xeł thida de seheghugha sêna seți nît’ha.
Ekwadi hot’a y’unâne desaze nask’edhe ekozi senasther xa
nânîyî nî thie k’è dânares’ith hot’â.
Nûneshi k’odene ejere detjuri xa ejere hanaldher ni ât’ila.
Eyer bekuwu hots’î tulu nask’edhe
eyer edachënaże bet’a ejere deljuri thetą eyi besânîniya.
Eyî edachënaže eîtth’i nagî ą nîstî hu
bek’e dânsa hots’î bek’e nahesgus.
Ekwat’u seansdher nî et’axa edachënaże naghiyes.
Bedzaraze hekker hot’â thu theyi astė.
Eku y’unâne honîgi hî ą be?ihe delk’osi bebethchêne thetą nes?î.
Eku hobeğha edesth’ak nįt’â asi
et’axa eîtth’i hili nahodher de
dene hîlchu hu tth’i dene darîth’i snî nît’â.
Eku sa duwe be?ihe delk’osi
se?î hutow tth’i țlaghe tai se?î de segha
holni hutow nesdhênus sa duwe.
Sîchchu losq nesdhên hot’â ta?e huyldû y’unâne nânjesa nî.
Nûneshi k’odene be ?edachënaże ghâ hunethidił nį.
Data File & Metadata Management

HOW 2

1. **Text Input**
   - make sure file is readable as .txt file

2. **Check**
   - check text for spelling consistency
   - add sentence punctuation
   - make copy of text file for inter-linearization or other analysis

3. **Save**
   - make sure text file contains no metadata
   - give appropriate file name (perhaps w/ metadata)
   - save text file as .txt file with UTF-8 encoding

4. **Place**
   - place .txt files in single corpus folder
Data File & Metadata Management

HOW

<table>
<thead>
<tr>
<th>File Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Goddard_001_M_nopunc_Ceremony for Otter.txt</td>
<td>A file containing ceremonial content for an Otter event.</td>
</tr>
<tr>
<td>Goddard_001_M_nopunc_Datsa Tthi.txt</td>
<td>A file containing information about Datsa Tthi.</td>
</tr>
<tr>
<td>Goddard_001_M_nopunc_Deneyuze Natser Heja.txt</td>
<td>A file containing information about Deneyuze Natser Heja.</td>
</tr>
<tr>
<td>Goddard_001_M_nopunc_Ebedanotilhe.txt</td>
<td>A file containing information about Ebedanotilhe.</td>
</tr>
<tr>
<td>Goddard_001_M_nopunc_His Legs Tremble.txt</td>
<td>A file containing content about someone's legs trembling.</td>
</tr>
<tr>
<td>JBlackman_Testimony_1893_Primrose Lake Hearings.txt</td>
<td>A file containing testimony from 1893 Primrose Lake Hearings.</td>
</tr>
<tr>
<td>JJanvier_001_I_Blué Quills Broad.txt</td>
<td>A file containing content related to blue quills.</td>
</tr>
<tr>
<td>JJanvier_001_I_Chedth Kats’alze.txt</td>
<td>A file containing content about Chedth Kats’alze.</td>
</tr>
<tr>
<td>JJanvier_001_I_Dene Xere Koe.txt</td>
<td>A file containing content about Dene Xere Koe.</td>
</tr>
<tr>
<td>JJanvier_001_I_Eire Thawe.txt</td>
<td>A file containing content about Eire Thawe.</td>
</tr>
<tr>
<td>JJanvier_001_I_Fairbanks.txt</td>
<td>A file containing content about Fairbanks.</td>
</tr>
<tr>
<td>JJanvier_001_I_Joe Louis.txt</td>
<td>A file containing content about Joe Louis.</td>
</tr>
<tr>
<td>JJanvier_001_I_Kasba Kuwe.txt</td>
<td>A file containing content about Kasba Kuwe.</td>
</tr>
<tr>
<td>JJanvier_001_I_Long Datsq Tthi.txt</td>
<td>A file containing content about Long Datsq Tthi.</td>
</tr>
<tr>
<td>JJanvier_001_I_Long_nopunc_Holdlie.txt</td>
<td>A file containing content about Holdlie.</td>
</tr>
<tr>
<td>JJanvier_001_I_nopunc_BigFish_Luwe Chok.txt</td>
<td>A file containing content about BigFish Luwe Chok.</td>
</tr>
<tr>
<td>JJanvier_001_I_nopunc_Lj Beghq Norlya.txt</td>
<td>A file containing content about Lj Beghq Norlya.</td>
</tr>
<tr>
<td>JJanvier_001_I_short_Datsa Tthi Naheht’yer.txt</td>
<td>A file containing content about Datsa Tthi Naheht’yer.</td>
</tr>
<tr>
<td>JJanvier_001_I_Xah Tuwe.txt</td>
<td>A file containing content about Xah Tuwe.</td>
</tr>
<tr>
<td>JJanvier_001_I_Dene Nahtghiya.txt</td>
<td>A file containing content about Dene Nahtghiya.</td>
</tr>
<tr>
<td>JJanvier_001_I_J_regret_Edelorensi.txt</td>
<td>A file containing content about regret and Edelorensi.</td>
</tr>
<tr>
<td>JJanvier_001_I_J_short_Holdlie.txt</td>
<td>A file containing content about Holdlie.</td>
</tr>
<tr>
<td>JJanvier_001_I_J_Xmas_Het’qoghe Yati.txt</td>
<td>A file containing content about Christmas and Het’qoghe Yati.</td>
</tr>
<tr>
<td>JJanvier_001_K_Fishing@Primrose_Xah Tuwe luwe Hoghq.txt</td>
<td>A file containing content about fishing at Primrose.</td>
</tr>
<tr>
<td>JJanvier001_K_ColdLakePeople_Luwe Chok Tuwe Hot’ine.txt</td>
<td>A file containing content about Cold Lake People and Luwe Chok Tuwe Hot’ine.</td>
</tr>
<tr>
<td>NMatchias_001_P_ChomokChomok.txt</td>
<td>A file containing content about ChomokChomok.</td>
</tr>
<tr>
<td>Petitot_001_B_Chap4_Nih Nahudle.txt</td>
<td>A file containing content about Chap4 Nih Nahudle.</td>
</tr>
<tr>
<td>Petitot_001_B_Batsune Yeneshi.txt</td>
<td>A file containing content about Batsune Yeneshi.</td>
</tr>
<tr>
<td>Petitot_001_B_Chap3_Einaheke.txt</td>
<td>A file containing content about Chap3 Einaheke.</td>
</tr>
<tr>
<td>Petitot_001_B_Chap5_Datsa.txt</td>
<td>A file containing content about Chap5 Datsa.</td>
</tr>
<tr>
<td>Petitot_001_B_nopunc_Deneyu.txt</td>
<td>A file containing content about nopunc Deneyu.</td>
</tr>
<tr>
<td>Petitot_AntoineUdahl_001_E_nopunc_Hochoghe.txt</td>
<td>A file containing content about AntoineUdahl E Hochoghe.</td>
</tr>
<tr>
<td>Petitot_Dene Yatie.txt</td>
<td>A file containing content about Dene Yatie.</td>
</tr>
<tr>
<td>Petitot_SmallNotebook_001_D_June 1863_SalTuCho_Datsq Dene Hodeluyg Dene à?át.txt</td>
<td>A file containing content about June 1863 SalTuCho Datsq Dene Hodeluyg Dene.</td>
</tr>
<tr>
<td>X_nopunc_Jackfish_Udahl Dene Hetheji.txt</td>
<td>A file containing content about Jackfish Udahl Dene Hetheji.</td>
</tr>
</tbody>
</table>

portion of the Dene Sųłiné corpus folder
Test-driving the Corpus

- classic corpus tools
  - word lists
  - dispersion
  - KWIC views
  - n-grams
  - collocates

- classic corpus applications
  - natural, example sentences
  - synonymy differentiation
  - sense disambiguation
  - genre/register/dialect differences
Test-driving the Corpus

<table>
<thead>
<tr>
<th>PLAINS CREE CORPUS FILE NAME</th>
<th>GENRE</th>
<th>SOURCE OF RECORDING</th>
<th>ORIGINAL MEDIA FILE NAME</th>
<th>transcribed (.xml file)</th>
<th>exported (.txt file)</th>
<th>length of recording</th>
<th># of word types</th>
<th># of word tokens</th>
<th>Type:Token ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>C_2F50_L&amp;S_AB-AB.txt</td>
<td>Conversation</td>
<td>CILLDI 2015</td>
<td></td>
<td></td>
<td></td>
<td>25:00:00</td>
<td>745</td>
<td>2006</td>
<td>0.37</td>
</tr>
<tr>
<td>C_2F50_T&amp;Z_AB-SK.txt</td>
<td>Conversation</td>
<td>CILLDI 2015</td>
<td></td>
<td></td>
<td></td>
<td>29:25:00</td>
<td>729</td>
<td>2099</td>
<td>0.35</td>
</tr>
<tr>
<td>C_2F50_R&amp;Z_SK.txt</td>
<td>Conversation</td>
<td>CILLDI 2015</td>
<td></td>
<td></td>
<td></td>
<td>36:53:00</td>
<td>1139</td>
<td>2902</td>
<td>0.39</td>
</tr>
<tr>
<td>N_1F70.MM_SK.txt</td>
<td>Narrative</td>
<td>YouTube</td>
<td></td>
<td></td>
<td></td>
<td>7:14</td>
<td>276</td>
<td>488</td>
<td>0.57</td>
</tr>
<tr>
<td>N_1M70_AR_SK.txt</td>
<td>Narrative</td>
<td>YouTube</td>
<td></td>
<td></td>
<td></td>
<td>5:50</td>
<td>247</td>
<td>424</td>
<td>0.58</td>
</tr>
<tr>
<td>N_2F60_3M60.SK.txt</td>
<td>Narrative</td>
<td>YouTube</td>
<td></td>
<td></td>
<td></td>
<td>3:30</td>
<td>116</td>
<td>163</td>
<td>0.71</td>
</tr>
<tr>
<td>S_Maskimocisya.txt</td>
<td>Story (Children’s)</td>
<td>B &amp; Z</td>
<td></td>
<td></td>
<td></td>
<td>print</td>
<td>92</td>
<td>143</td>
<td>0.64</td>
</tr>
<tr>
<td>S_Mosīm.txt</td>
<td>Story (Children’s)</td>
<td>B &amp; Z</td>
<td></td>
<td></td>
<td></td>
<td>print</td>
<td>130</td>
<td>254</td>
<td>0.51</td>
</tr>
<tr>
<td>S_Osāwistikwān.txt</td>
<td>Story (Children’s)</td>
<td>CILLDI 2015</td>
<td></td>
<td></td>
<td></td>
<td>print</td>
<td>139</td>
<td>251</td>
<td>0.55</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2917</td>
<td>8730</td>
<td></td>
</tr>
</tbody>
</table>

portion of the Plains Cree metadata file
Test-driving the Corpus AntConc interface
Test-driving the Corpus word lists ALPHABETICAL

kiskinoham (VTI)
‘s/he marks s.th.’
‘s/he guides s.th.’
‘s/he points out s.th.’
Test-driving the Corpus word lists

mâka  ‘but, then, still’
mâna  ‘usually, always’
mêtoni  ‘very, really’
Test-driving the Corpus word lists by frequency

1. aya ‘ah, well, um, hmm’
2. êkwa ‘and, also; then, now’
7. anima ‘that, the fact that’
14. êsa ‘apparently, evidently’
Test-driving the Corpus dispersion

aya
‘ah, well, um, hmm’
Test-driving the Corpus dispersion

<table>
<thead>
<tr>
<th>Corpus Files</th>
<th>File View</th>
<th>Clusters/n-Grams</th>
<th>Collocates</th>
<th>Word List</th>
<th>Keyword List</th>
</tr>
</thead>
<tbody>
<tr>
<td>C_2F50_L5S_4B.txt</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C_2F50_L5B_Z_SK.txt</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C_2F50_T5Z_4B_SK.txt</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N_1FT0_W_M_SK.txt</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N_1M170_A4_SK.txt</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N_2F50_3M50_SK.txt</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S_Maskemocie.txt</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S_Mosom.txt</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S_OsAmnlatikwàn.txt</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Concordance Hits**

| HIT FILE: 1 | FILE: C_2F50_L5S_4B.txt | No. of Hits = 19 | File Length (in chars) = 16016 |
| HIT FILE: 2 | FILE: C_2F50_L5B_Z_SK.txt | No. of Hits = 15 | File Length (in chars) = 24015 |
| HIT FILE: 3 | FILE: C_2F50_T5Z_4B_SK.txt | No. of Hits = 5 | File Length (in chars) = 15648 |
| HIT FILE: 4 | FILE: N_1FT0_W_M_SK.txt | No. of Hits = 3 | File Length (in chars) = 3841 |
| HIT FILE: 5 | FILE: N_1M170_A4_SK.txt | No. of Hits = 10 | File Length (in chars) = 3817 |
| HIT FILE: 6 | FILE: N_2F50_3M50_SK.txt | No. of Hits = 2 | File Length (in chars) = 1481 |
| HIT FILE: 7 | FILE: S_Maskemocie.txt | No. of Hits = 3 | File Length (in chars) = 1319 |
| HIT FILE: 8 | FILE: S_Mosom.txt | No. of Hits = 1 | File Length (in chars) = 1092 |
| HIT FILE: 9 | FILE: S_OsAmnlatikwàn.txt | No. of Hits = 1 | File Length (in chars) = 1001 |

**Search Term**

- kîkway
  - ‘something’
  - ‘anything’
Test-driving the Corpus

**KWIC views**

kíkway
‘something’
‘anything’
Test-driving the Corpus

KWIC views

kîkway
‘something’
‘anything’
Test-driving the Corpus n-grams

ëkwa aya    ‘and then, um’
ëkwa màna    ‘and then, usually’
ëkwa mîna    ‘and then, again’
ëkwa namôya  ‘and then, not’
Test-driving the Corpus **collocates**

<table>
<thead>
<tr>
<th>Rank</th>
<th>Type</th>
<th>Freq</th>
<th>Freq(L)</th>
<th>Freq(R)</th>
<th>Stat</th>
<th>Collocate</th>
<th>Total No. of Collocate Tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>9</td>
<td>2</td>
<td>7</td>
<td></td>
<td>5.47216</td>
<td>kikway</td>
<td>13</td>
</tr>
<tr>
<td>2</td>
<td>7</td>
<td>3</td>
<td>4</td>
<td></td>
<td>4.67030</td>
<td>ewako</td>
<td>12</td>
</tr>
<tr>
<td>3</td>
<td>6</td>
<td>4</td>
<td>2</td>
<td></td>
<td>4.26284</td>
<td>ekoto</td>
<td>11</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>1</td>
<td>3</td>
<td></td>
<td>4.03513</td>
<td>ndi'o</td>
<td>10</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td></td>
<td>3.75861</td>
<td>ndi'o</td>
<td>9</td>
</tr>
<tr>
<td>6</td>
<td>7</td>
<td>2</td>
<td>5</td>
<td></td>
<td>3.43764</td>
<td>owa'n</td>
<td>8</td>
</tr>
<tr>
<td>7</td>
<td>9</td>
<td>6</td>
<td>3</td>
<td></td>
<td>3.39480</td>
<td>ndi'o</td>
<td>7</td>
</tr>
<tr>
<td>8</td>
<td>16</td>
<td>8</td>
<td>8</td>
<td></td>
<td>3.32689</td>
<td>aya</td>
<td>6</td>
</tr>
<tr>
<td>9</td>
<td>13</td>
<td>10</td>
<td>3</td>
<td></td>
<td>3.26326</td>
<td>ekwa</td>
<td>5</td>
</tr>
<tr>
<td>10</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td></td>
<td>3.14048</td>
<td>antla</td>
<td>4</td>
</tr>
<tr>
<td>11</td>
<td>4</td>
<td>3</td>
<td>1</td>
<td></td>
<td>3.03513</td>
<td>ndi'o</td>
<td>4</td>
</tr>
</tbody>
</table>

**mistahi kikway** ‘a lot of something’
Test-driving the Corpus

REGEX searches

Cree locative suffix -ihk, -ohk, -ahk

\w+(i|o|a)hk\s
Test-driving the Corpus

REGEX searches

word-internal string
(-)nitawi- ‘go and’

\w+nitawi\w+
Test-driving the Corpus

other applications

concordance lines

sense disambiguation

synonym differences
In the world of corpus linguistics, bigger may be better....

- but sampling skews are part of the reality of (endangered) language documentation

- do as much as you can with the language samples you’ve got!
  (so you’re inspired to collect more samples more broadly from more usage situations and more speakers)

- a small, untagged, & unbalanced corpus can still yield tremendous insights into the structure, meaning, & use of a language

Your community probably has all the tools and personnel you need to start building and using a corpus.
In the world of language documentation & revitalization...

- making samples of language accessible and reusable for multiple purposes and multiple users is key

- taking the language as it comes (not based on translation, elicitation, or someone else’s analysis) is the best starting point for good documentation

- the creation and maintenance of a language corpus can involve a wide variety of community members with differing skills & interests

You don’t need an outside linguist or programmer!
So, go build a corpus...

...and feed your language community!
Ish nish
Masi cho
Hay-hay
Mahalo nui

sally.rice@ualberta.ca  dthunder@ualberta.ca
References

- Johnson, H. 2008. Language documentation and archiving, or how to build a better corpus (ms.)

Special thanks to Brenda Morency and Eugene Alexis!