Bootstrapping a Neural Morphological Analyzer for St. Lawrence Island Yupik Nouns from a Finite-State Transducer

Lane Schwartz ¹  Emily Chen ¹  Sylvia Schreiner ²  Benjamin Hunt ²

¹University of Illinois Urbana-Champaign
²George Mason University

February 27, 2019
INTRODUCTION

► About St. Lawrence Island Yupik
  * Member of the Inuit-Yupik language family and spoken on St. Lawrence Island, AK
  * ∼1000 L1 speakers remaining
  * Endangered and low-resource

► Developing computational resources for Yupik to assist with the revitalization effort

► Introduce a neural morphological analyzer for Yupik nouns today
YUPIK MORPHOLOGY

- Yupik is **polysynthetic**, allowing for morphologically-complex words

(1) **mangteghaghllangllaghyugtukut**

mangteghagh- -ghllag- -ngllagh- -yug- -tu- -kut
house- -big- -build- -want.to- -INTR.IND- -1PL

‘We want to build a big house’

- Yupik words typically adhere to the following template:

Root + 0-7 Derivational Morpheme(s) + Inflectional Morphemes + (Enclitic)
YUPIK MORPHOLOGY

- Yupik is **polysynthetic**, allowing for morphologically-complex words

(1) **mangteghaghllangllaghyugtukut**

- mangteghagh- -ghllag- -ngllagh- -yug- -tu- -kut
- house- -big- -build- -want.to- -INTR.IND- -1PL

‘*We want to build a big house’*

- Yupik words typically adhere to the following template:

  **Root** + 0-7 Derivational Morpheme(s) + Inflectional Morphemes + (Enclitic)
YUPIK MORPHOLOGY

- Yupik is **polysynthetic**, allowing for morphologically-complex words

(1) **mangteghaghllangllaghyugtukut**

*Root: mangteghagh-*,
*Derivational Morphemes: ghllag-, ngllagh-, yug-*,
*Inflectional Morphemes: tu-*,
* Pronoun: -kut*.

*House-*
*Derivational Morphemes: big-*,
*Inflectional Morphemes: build-, want.to-*,
*Indicative: INTR.IND-*,
* Pronoun: -1PL (we, first person, plural)*

‘We want to build a big house’

- Yupik words typically adhere to the following template:

Root + 0-7 Derivational Morpheme(s) + Inflectional Morphemes + (Enclitic)
YUPIK MORPHOLOGY

- Yupik is **polysynthetic**, allowing for morphologically-complex words

(1) **mangteghaghllangllaghyugtukut**
  
  mangteghagh-   -ghllag-   -ngllagh-   -yug-   -tu-   -kut  
  house-   -big-   -build-   -want.to-   -INTR.IND-   -1PL

  ‘*We want to build a big house*’

- Yupik words typically adhere to the following template:

Root + 0-7 Derivational Morpheme(s) + Inflectional Morphemes + (Enclitic)
YUPIK MORPHOPHONOLOGY

- Yupik also exhibits morphophonological properties during suffixation of morphemes

\[(1) \text{mangteghaghllangllaghyugtukut}\]

mangteghagh- -ghllag- -ngllagh- -yug- -tu- -kut
house- -big- -build- -want.to- -INTR.IND- -1PL

‘We want to build a big house’

TAKEAWAYS

- Morphophonology does occur and is a critical aspect of Yupik morphology
- It complicates the affixation of morphemes in Yupik, blurring the boundaries that otherwise exist between each constituent morpheme
Morphological analysis is the parsing of a given word (the surface form) into its constituent morphemes (the underlying form).

Surface: mangteghaghllangllaghyugtukut

Underlying: mangteghagh-ghllag-ngllagh-yug-INTR.IND-1PL

Developing a morphological analyzer for Yupik is challenging since its morphophonology may obscure morpheme boundaries.
FIRST ATTEMPT: Implemented a finite-state analyzer for Yupik (Chen & Schwartz, 2018) using the Foma finite-state toolkit (Hulden, 2009)

Evaluated by calculating its coverage

coverage = \frac{\text{Number of Words Analyzed}}{\text{Number of Words in Text}}

<table>
<thead>
<tr>
<th>Text</th>
<th>Coverage (%)</th>
<th>Token Count</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Tokens; Types</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>98.24</td>
<td>97.87</td>
</tr>
<tr>
<td>2</td>
<td>79.10</td>
<td>70.62</td>
</tr>
<tr>
<td>3</td>
<td>77.14</td>
<td>68.87</td>
</tr>
<tr>
<td>4</td>
<td>76.98</td>
<td>68.32</td>
</tr>
<tr>
<td>5</td>
<td>84.08</td>
<td>73.45</td>
</tr>
<tr>
<td>6</td>
<td>76.64</td>
<td>70.86</td>
</tr>
<tr>
<td>7</td>
<td>75.42</td>
<td>72.62</td>
</tr>
<tr>
<td>8</td>
<td>77.71</td>
<td>75.19</td>
</tr>
<tr>
<td>Average</td>
<td>80.57</td>
<td>74.73</td>
</tr>
</tbody>
</table>
Attempted to extend coverage of the finite-state analyzer through fieldwork

* Managed to elicit previously undocumented lexical items and grammatical constructions
* But method was highly dependent on speaker availability and knowledge
* Was not an optimal use of time and resources

ALTERNATIVE METHOD (Micher, 2017; Moeller et al., 2018)

1. Recast morphological analysis as a machine translation task
2. Use the finite-state analyzer to mass generate surface form-glossed form pairs
3. Train the neural morphological analyzer on this generated dataset
MORPHOLOGICAL ANALYSIS AS MACHINE TRANSLATION

Morphological analysis can be recast as a machine translation task:

\[
mangteghaq \\
\downarrow \\
mangteghagh[N][ABS][SG]
\]

Generated dataset was subsequently tokenized as follows:

* by **character**

\[
\text{m a n g t e g h a q} \\
\text{m a n g t e g h a g h [N] [ABS] [SG]}
\]

* by **grapheme**

\[
\text{m a n g t e g h a q} \\
\text{m a n g t e g h a g h [N] [ABS] [SG]}
\]
**DATASET**

- **OBJECTIVE**: Develop a neural morphological analyzer for analyzing inflected Yupik nouns with no derivational morphology

- **TRAINING DATA**: A parallel dataset consisting of every inflected noun and its underlying form
  
  * Paired every Yupik noun root with every nominal inflectional suffix

<table>
<thead>
<tr>
<th>Noun Root</th>
<th>Inflectional Suffix</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Case</td>
<td>Number</td>
</tr>
<tr>
<td>3873</td>
<td>7</td>
<td>3</td>
</tr>
<tr>
<td>3873</td>
<td>7</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Underlying Form</td>
<td></td>
<td></td>
</tr>
<tr>
<td>----------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>mangteghagh[N][ABS][SG]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>mangteghagh[N][ABS][PL]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>mangteghagh[N][ABS][DU]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>mangteghagh[N][ABS][SG][3SGPOSS]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>mangteghagh[N][ABS][SG][3PLPOSS]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>mangteghagh[N][ABS][SG][3DUPOSS]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>mangteghagh[N][VIA][DU][4SGPOSS]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>mangteghagh[N][VIA][DU][4PLPOSS]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>mangteghagh[N][VIA][DU][4DUPOSS]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Underlying Form</td>
<td>Surface Form</td>
<td></td>
</tr>
<tr>
<td>---------------------------------------------</td>
<td>----------------</td>
<td></td>
</tr>
<tr>
<td>mangteghagh[N][ABS][SG]</td>
<td>mangteghaq</td>
<td></td>
</tr>
<tr>
<td>mangteghagh[N][ABS][PL]</td>
<td>mangteghaat</td>
<td></td>
</tr>
<tr>
<td>mangteghagh[N][ABS][DU]</td>
<td>mangteghaak</td>
<td></td>
</tr>
<tr>
<td>mangteghagh[N][ABS][SG][3SGPOSS]</td>
<td>mangteghaa</td>
<td></td>
</tr>
<tr>
<td>mangteghagh[N][ABS][SG][3PLPOSS]</td>
<td>mangteghaat</td>
<td></td>
</tr>
<tr>
<td>mangteghagh[N][ABS][SG][3DUPOSS]</td>
<td>mangteghaak</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td></td>
</tr>
<tr>
<td>mangteghagh[N][VIA][DU][4SGPOSS]</td>
<td>mangteghagmikun</td>
<td></td>
</tr>
<tr>
<td>mangteghagh[N][VIA][DU][4PLPOSS]</td>
<td>mangteghagmegteggun</td>
<td></td>
</tr>
<tr>
<td>mangteghagh[N][VIA][DU][4DUPOSS]</td>
<td>mangteghagmegtegnegun</td>
<td></td>
</tr>
</tbody>
</table>
IMPLEMENTED THE NEURAL ANALYZER IN MarianNMT (Junczys-Dowmunt et al., 2018)

- encoder-decoder model
- recurrent
- bidirectional
- attentional

* Implemented a shallow model with one hidden layer

* Randomly partitioned the 1,057,329-item dataset as follows:
  - TRAINING SET: 80%
  - VALIDATION SET: 10%
  - TEST SET: 10%

* Tokenized the partitioned datasets by character

* Achieved 100% coverage and 59.67% accuracy
DEBUGGING

- Encountered an issue with **case syncretism**:

  (2a)  
  ayveghet
  ayvegh- et
  walrus- ABS.PL
  ‘walruses’

  (2b)  
  ayveghet
  ayvegh- et
  walrus- ERG.PL
  ‘of walruses’

- Checked if the surface form of the neural analyzer’s output matched the surface form of the test set’s output

<table>
<thead>
<tr>
<th></th>
<th>Output</th>
<th>Surface</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neural Analyzer</td>
<td>ayvegh[N][ABS][PL]</td>
<td>ayveghat</td>
</tr>
<tr>
<td>Test Set</td>
<td>ayvegh[N][ERG][PL]</td>
<td>ayveghat</td>
</tr>
</tbody>
</table>
DEBUGGING

- Encountered an issue with **case syncretism**:

(2a) **ayveghet**
    ayvegh- -et
    walrus- **ABS.PL**
    ‘walruses’

(2b) **ayveghet**
    ayvegh- -et
    walrus- **ERG.PL**
    ‘of walruses’

- Checked if the surface form of the neural analyzer’s output matched the surface form of the test set’s output

<table>
<thead>
<tr>
<th>Output</th>
<th>Surface</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Neural Analyzer</strong></td>
<td>ayvegh[N][ABS][PL]</td>
</tr>
<tr>
<td><strong>Test Set</strong></td>
<td>ayvegh[N][LOC][PL]</td>
</tr>
</tbody>
</table>
DEBUGGING

- Encountered an issue with **case syncretism**:

  (2a)  
  ayveghet  
  ayvegh- -et  
  walrus- -ABS.PL  
  ‘walruses’

  (2b)  
  ayveghet  
  ayvegh- -et  
  walrus- -ERG.PL  
  ‘of walruses’

- Checked if the surface form of the neural analyzer’s output matched the surface form of the test set’s output

- Achieved **100% coverage** and **99.90% accuracy**
ADDITIONAL EXPERIMENTS

- Trained four additional models, experimenting with the tokenization scheme and depth of the model.

- All else remained the same as the model from the initial run.

- Results

<table>
<thead>
<tr>
<th></th>
<th>character</th>
<th>grapheme</th>
</tr>
</thead>
<tbody>
<tr>
<td>shallow</td>
<td>99.87%</td>
<td>99.90%</td>
</tr>
<tr>
<td>deep</td>
<td>99.95%</td>
<td>99.96%</td>
</tr>
</tbody>
</table>
EVALUATION OF THE NEURAL ANALYZER

▶ EVALUATION OBJECTIVES

1. Evaluate the performance of the neural analyzers on a blind test set

2. Contrast the performance of the neural analyzer with the performance of the finite-state analyzer

▶ Supplemented the finite-state analyzer with a guesser module

* Permits the analyzer to hypothesize possible roots
* All guesses adhere to Yupik phonotactics and syllable structure
**BLIND TEST SET & RESULTS**

- **BLIND TEST SET**: *Mrs. Della Waghiyi’s St. Lawrence Island Yupik Texts With Grammatical Analysis by Kayo Nagai* (Waghiyi & Nagai, 2001)

  * Identified **344 inflected nouns with no derivational morphology**

- **Types**

<table>
<thead>
<tr>
<th></th>
<th>Coverage (%)</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>FST (No Guesser)</td>
<td>85.78</td>
<td>78.90</td>
</tr>
<tr>
<td>FST (w/Guesser)</td>
<td>100</td>
<td>84.86</td>
</tr>
<tr>
<td>Neural</td>
<td>100</td>
<td><strong>92.20</strong></td>
</tr>
</tbody>
</table>

- **Tokens**

<table>
<thead>
<tr>
<th></th>
<th>Coverage (%)</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>FST (No Guesser)</td>
<td>85.96</td>
<td>79.82</td>
</tr>
<tr>
<td>FST (w/Guesser)</td>
<td>100</td>
<td>84.50</td>
</tr>
<tr>
<td>Neural</td>
<td>100</td>
<td><strong>91.81</strong></td>
</tr>
</tbody>
</table>
CAPACITY TO GENERALIZE

- The neural analyzer fared better on OOV or unattested roots:

<table>
<thead>
<tr>
<th>OOV Root</th>
<th>FST</th>
<th>NN</th>
</tr>
</thead>
<tbody>
<tr>
<td>aghnasinghagh</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>aghveghniigh</td>
<td>–</td>
<td>✓</td>
</tr>
<tr>
<td>akughvigaggh</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>qikmiraagh</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>sakara</td>
<td>✓</td>
<td>–</td>
</tr>
<tr>
<td>sanaghte</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>tangiqagh</td>
<td>–</td>
<td>✓</td>
</tr>
</tbody>
</table>

- The neural analyzer also fared better on spelling variants:

<table>
<thead>
<tr>
<th>Root Variant</th>
<th>FST</th>
<th>NN</th>
</tr>
</thead>
<tbody>
<tr>
<td>melqighagh</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>piitesiiighagh</td>
<td>–</td>
<td>✓</td>
</tr>
<tr>
<td>uqfiillegagh</td>
<td>–</td>
<td>✓</td>
</tr>
<tr>
<td>*ukusumun</td>
<td>–</td>
<td>✓</td>
</tr>
</tbody>
</table>
CONCLUSION

- Introduced a neural morphological analyzer for Yupik nouns with no derivational morphology

- Showed how a high-performing morphological analyzer can be bootstrapped from an existing finite-state analyzer

- **Implications** for . . .
  - Other Low-Resource Languages
  - Fieldwork

- **Future Work**
  - Select a tokenization scheme and model depth
  - Consider handling of syncretic items
  - Implement a neural analyzer for the full Yupik lexicon
Thank you!

Questions?
REFERENCES


