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

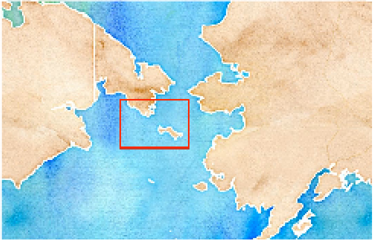
Lane Schwartz [1] Emily Chen [1] Sylvia Schreiner [2] Benjamin Hunt [2]

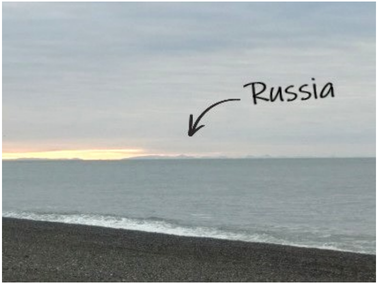
[1]University of Illinois Urbana-Champaign

[2]George Mason University

February 27, 2019

- About St. Lawrence Island Yupik
  - Member of the Inuit-Yupik language family and spoken on St. Lawrence Island, AK
  - $\sim$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 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 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 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 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 also exhibits morphophonological properties during suffixation of morphemes

(1) **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** $= \frac{\text{Number of Words Analyzed}}{\text{Number of Words in Text}}$

Text	**Coverage** (%) Tokens	Types	**Token Count**
1	98.24	97.87	795
2	79.10	70.62	6859
3	77.14	68.87	11,926
4	76.98	68.32	12,982
5	84.08	73.45	15,766
6	76.64	70.86	4357
7	75.42	72.62	5358
8	77.71	75.19	5731
Average	**80.57**	**74.73**	63,774

- 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 can be recast as a machine translation task:

mangteghaq

↓

mangteghagh[N][ABS][SG]

- Generated dataset was subsequently tokenized as follows:
  - by **character**

    m a n g t e g h a q
    m a n g t e g h a g h [N] [ABS] [SG]

  - by **grapheme**

    m a ng t e gh a q
    m a ng t e gh a gh [N] [ABS] [SG]

- **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

Noun Root	Inflectional Suffix				TOTAL
	Case	**Number**	**Possession**		
			Person	*Number*	
3873	7	3	–	–	81,333
3873	7	3	4	3	975,996
					1,057,329

Underlying Form
mangteghagh[N][ABS][SG]
mangteghagh[N][ABS][PL]
mangteghagh[N][ABS][DU]
mangteghagh[N][ABS][SG][3SGPOSS]
mangteghagh[N][ABS][SG][3PLPOSS]
mangteghagh[N][ABS][SG][3DUPOSS]
⋮
⋮
⋮
mangteghagh[N][VIA][DU][4SGPOSS]
mangteghagh[N][VIA][DU][4PLPOSS]
mangteghagh[N][VIA][DU][4DUPOSS]

Underlying Form	Surface Form
mangteghagh[N][ABS][SG]	mangteghaq
mangteghagh[N][ABS][PL]	mangteghaat
mangteghagh[N][ABS][DU]	mangteghaak
mangteghagh[N][ABS][SG][3SGPOSS]	mangteghaa
mangteghagh[N][ABS][SG][3PLPOSS]	mangteghaat
mangteghagh[N][ABS][SG][3DUPOSS]	mangteghaak
⋮	
⋮	
⋮	
mangteghagh[N][VIA][DU][4SGPOSS]	mangteghagmikun
mangteghagh[N][VIA][DU][4PLPOSS]	mangteghagmegteggun
mangteghagh[N][VIA][DU][4DUPOSS]	mangteghagmegtegnegun

- Implemented the neural analyzer in MarianNMT (Junczys-Dowmunt et al., 2018)
  - encoder-decoder model
  - recurrent
  - bidirectional
  - attentional
- **INITIAL RUN**
  - 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

- 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

	Output	**Surface**
Neural Analyzer	ayvegh[N][ABS][PL]	ayveghat
Test Set	ayvegh[N][ERG][PL]	ayveghat
		✓

- 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

	Output	**Surface**
Neural Analyzer	ayvegh[N][ABS][PL]	ayveghat
Test Set	ayvegh[N][LOC][PL]	ayveghni
		✗

- 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

- 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

	character	**grapheme**
shallow	99.87%	99.90%
deep	99.95%	**99.96%**

- **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**: *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

	Coverage (%)	**Accuracy** (%)
FST (No Guesser)	85.78	78.90
FST (w/Guesser)	100	84.86
Neural	100	**92.20**

- Tokens

	Coverage (%)	**Accuracy** (%)
FST (No Guesser)	85.96	79.82
FST (w/Guesser)	100	84.50
Neural	100	**91.81**

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

OOV Root	FST	NN
aghnasinghagh	–	–
aghveghniigh	–	✓
akughvigagh	✓	✓
qikmiraagh	–	–
sakara	✓	–
sanaghte	–	–
tangiqagh	–	✓

- The neural analyzer also fared better on spelling variants:

Root Variant	FST	NN
melq**i**ghagh	✓	✓
piites**ii**ghagh	–	✓
uqf**ii**lleghagh	–	✓
*uk**u**sumun	–	✓

- 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

Apassingok, A., Uglowook, J., Koonooka, L., and Tennant, E., editors (1993). *Kallagneghet / Drumbeats*. Bering Strait School District, Unalakleet, Alaska.

Apassingok, A., Uglowook, J., Koonooka, L., and Tennant, E., editors (1994). *Akiingqwaghneghet / Echoes*. Bering Strait School District, Unalakleet, Alaska.

Apassingok, A., Uglowook, J., Koonooka, L., and Tennant, E., editors (1995). *Suluwet / Whisperings*. Bering Strait School District, Unalakleet, Alaska.

Apassingok, A., Walunga, W., and Tennant, E., editors (1985). *Sivuqam Nangaghnegha — Siivanllemta Ungipaqellghat / Lore of St. Lawrence Island — Echoes of our Eskimo Elders*, volume 1: Gambell. Bering Strait School District, Unalakleet, Alaska.

Apassingok, A., Walunga, W., and Tennant, E., editors (1987). *Sivuqam Nangaghnegha — Siivanllemta Ungipaqellghat / Lore of St. Lawrence Island — Echoes of our Eskimo Elders*, volume 2: Savoonga. Bering Strait School District, Unalakleet, Alaska.

Apassingok, A., Walunga, W., and Tennant, E., editors (1989). *Sivuqam Nangaghnegha — Siivanllemta Ungipaqellghat / Lore of St. Lawrence Island — Echoes of our Eskimo Elders*, volume 3: Southwest Cape. Bering Strait School District, Unalakleet, Alaska.

Badten, L. W., Kaneshiro, V. O., Oovi, M., and Koonooka, C. (2008). *St. Lawrence Island / Siberian Yupik Eskimo Dictionary*. Alaska Native Language Center, University of Alaska Fairbanks.

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Chen, E. and Schwartz, L. (2018). A morphological analyzer for St. Lawrence Island / Central Siberian Yupik. *In Proceedings of the 11th Language Resources and Evaluation Conference*, Miyazaki, Japan.

Hulden, M. (2009). Foma: A finite-state compiler and library. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 29–32. Association for Computational Linguistics.

Jacobson, S. A. (2001). *A Practical Grammar of the St. Lawrence Island / Siberian Yupik Eskimo Language, Preliminary Edition*. Alaska Native Language Center, Fairbanks, Alaska, 2nd edition.

Junczys-Dowmunt, M., Grundkiewicz, R., Dwojak, T., Hoang, H., Heafield, K., Neckermann, T., Seide, F., Germann, U., Fikri Aji, A., Bogoychev, N., Martins, A. F. T., and Birch, A. (2018). Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

Koonooka, C. P. (2003). *Ungipaghaghlanga: Let Me Tell You A Story*. Alaska Native Language Center.

Koonooka, C. P. (2005). Yupik language instruction in Gambell (St. Lawrence Island, Alaska). *Etudes/Inuit/Studies*, 29(1/2):251–266.

Micher, J. (2017). Improving coverage of an Inuktitut morphological analyzer using a segmental recurrent neural network. In *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages pp. 101–106, Honolulu. Association for Computational Linguistics.

Moeller, S., Kazeminejad, G., Cowell, A., and Hulden, M. (2018). A neural morphological analyzer for arapaho verbs learned from a finite state transducer. In *Proceedings of the Workshop on Computational Modeling of Polysynthetic Languages*, Santa Fe, New Mexico. Association for Computational Linguistics.

Morgounova, D. (2007). Language, identities and ideologies of the past and present Chukotka. *Etudes/Inuit/Studies*, 31(1-2):183–200.

Nagai, K. (2001). *Mrs. Della Waghiyi's St. Lawrence Island Yupik Texts with Grammatical Analysis. Number A2-006 in Endangered Languages of the Pacific Rim*. Nakanishi Printing, Kyoto, Japan.

Schuster, M. and Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681.