Bidirectional Leveraging of Computational Morphology and Linguistic Fieldwork

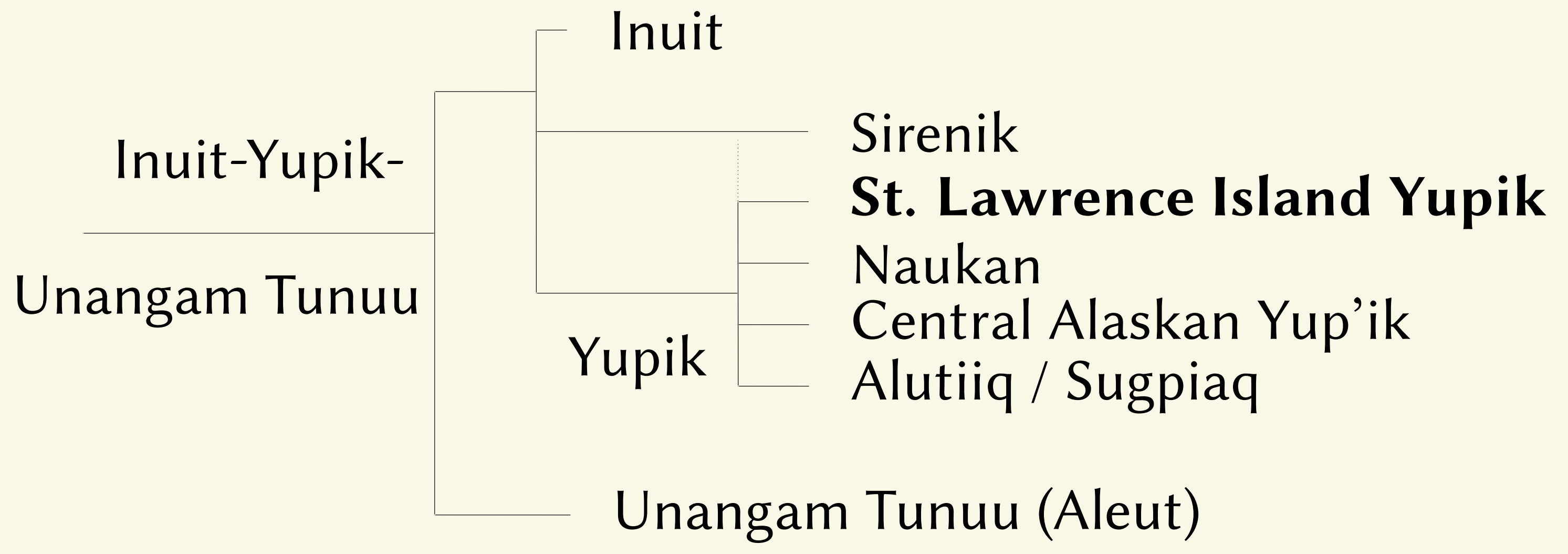
Lane Schwartz [1] Sylvia L.R. Schreiner [2] Emily Chen [1] Benjamin Hunt [2]

[1]University of Illinois Urbana-Champaign [2]George Mason University

Overview

St. Lawrence Island Yupik is an endangered language of the Bering Strait region. As a polysynthetic language, the availability of a high-coverage morphological analyzer is a prerequisite for the development of other computational resources for Yupik. Our existing morphological analyzer (Chen & Schwartz, 2018) failed to provide an analysis for approximately 25% of word types in our digitized corpus of Yupik texts. The questions raised from these failures guided subsequent fieldwork sessions, where we successfully identified previously undescribed lexical, morphological, and phonological processes in Yupik. This led to increased coverage of the morphological analyzer, resulting in a ***virtuous cycle*** that jointly leverages computational morphology and linguistic fieldwork.

St. Lawrence Island Yupik



Digitize Legacy Resources

Numerous Yupik-language texts were developed in the Soviet Union in the early 20th century and in Alaska in the mid- to late-20th century. One project goal is the digitization of these texts. To date, we have digitized Apassingok et al. (1985, 1987, 1989, 1993, 1994, 1995) and Koonooka (2003).

Yupik Grammar Overview

- Polysynthetic
- Ergative-absolutive
- 4 persons, 3 numbers
- Fairly free word order
- ~500 particles
- Extensive system of demonstratives
- ~600+ derivational suffixes
- General structure (inflected verb):

Root + Derivation + NEG + TMMA + Infl + Encl

Language Documentation & Analysis

A major goal of our fieldwork is the documentation of Yupik phonology, morphology, and syntax beyond that described by Krauss (1975) & Jacobson (2001). We hope this will be of use for developing modern pedagogical materials for Yupik language instruction and immersion programs.

	L1 Yupik Speakers	Yupik Population
Mainland Russia	<200	800
St. Lawrence Island	500—700	1300
Mainland Alaska	<200	400
Total	800—900	2400—2500

Fieldwork on St. Lawrence Island

- Semi-naturalistic production
- Targeted elicitation of morphosyntactic / semantic phenomena and analyzer errors
- Detailed positional and semantic work with derivational morphology
- Translation of Yupik texts into English
- Expansion of current lexicon

Computational Tool Development

We view computational tool development as integral to language documentation and revitalization. To date, we have developed a suite of web-based orthographic utilities (Schwartz and Chen, 2017), a finite-state morphological analyzer (Chen and Schwartz, 2018), a preliminary neural morphological analyzer (Schwartz et al., 2019), and an electronic dictionary (Hunt et al., 2019).

Morphological Analysis in the Field

- Finite-state analyzer implements Yupik grammar of Jacobson (2001) using `foma` (Hulden, 2009)
- User provides Yupik surface form.
  Potential morphological analyses are returned:

```
apply up> mangteghaghllagek
mangteghagh–ghllag[N→N][N][Abs][Unpd][Du]
mangteghagh–ghllag[N→N][N][Rel][Unpd][Du]
```

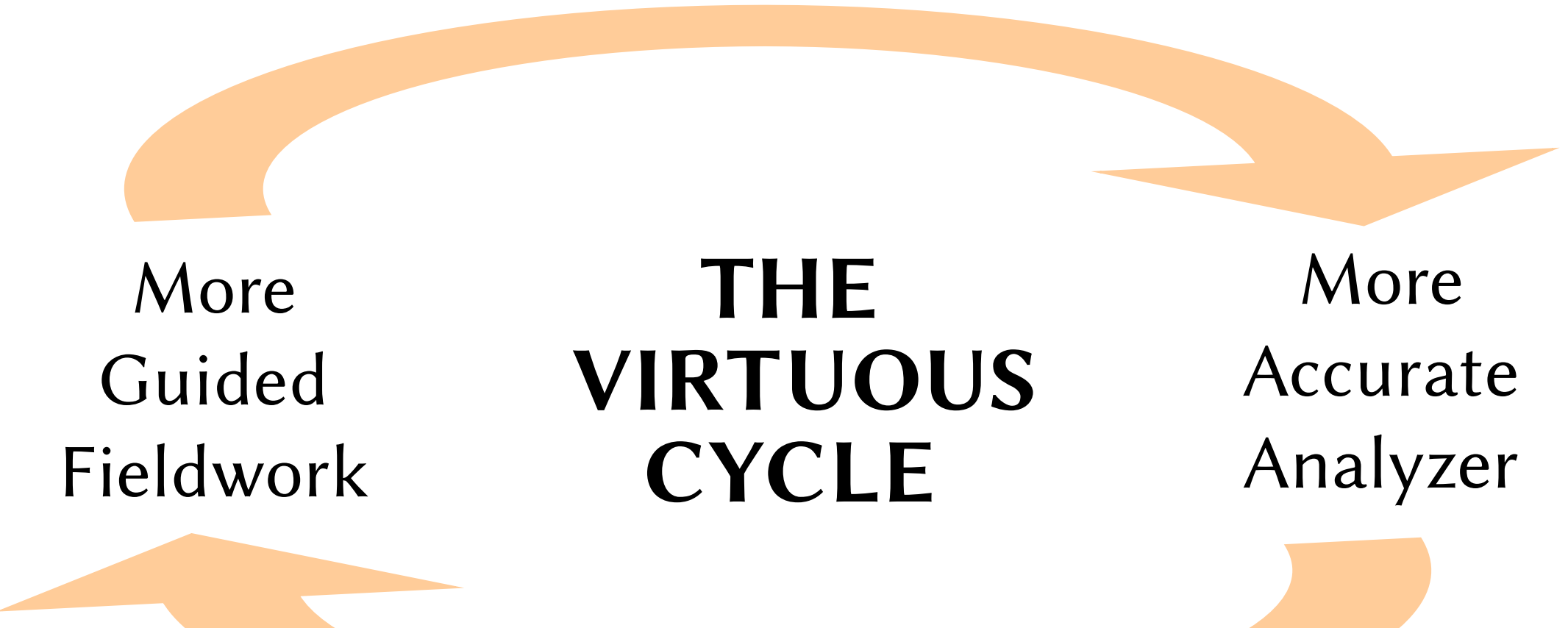
- User provides Yupik underlying form.
  Corresponding Yupik surface form is returned:

```
apply down> mangteghagh–ghllag[N→N][N][Abs][Unpd][Du]
mangteghaghllagek
```

Processes and Insights from the Field: Establishing the *Virtuous Cycle*

The Virtuous Cycle

- **SCENARIO 1**: Analyzer fails to analyze a word or produce the known correct analysis
  - *Hypothesis*: Word may involve linguistic phenomena that are currently undocumented or not well-documented
  - *Solution*: Inquire about the word with a speaker; adjust analyzer or lexicon as appropriate
- Fieldwork Elicitation
  ↓
  Larger Corpus
  ↓
  More Accurate Morphological Analyzer



Examples

- *aatqus*, *aghnas*, *akughvigaas*, etc.
  Previously undocumented phonological (and orthographic) process applying across word boundaries: word-final /t/ → /s/ when followed by word-initial /t/
- *allgeqestaamaan* - Previously undocumented derivational suffix, -*kestamaan* ("in the time of")

The Virtuous Cycle (CONT)

- **SCENARIO 2**: Elicitor successfully analyzes a word with the analyzer
  - *Result 1*: One Analysis → Follow-up with speaker to confirm correctness
  - *Result 2*: Multiple Analyses → Follow-up with speaker to determine the correct analysis
- More Accurate Morphological Analyzer
  ↓
  More Efficient Analysis of New Data
  ↓
  More Guided Future Elicitations

Implications

- The *virtuous cycle* benefits all sides of the research process:
  - **Computational Linguistics**
    - Analyzer is improved more quickly and more accurately
    - Analyzer facilitates development of computational resources for the community
  - **Language Documentation**
    - Elicitation/documentation is expedited with a better-performing analyzer
    - Documentation is improved as the analyzer identifies gaps in existing descriptions
  - **Related Languages**
    - Other languages in the family may benefit from the improved documentation of Yupik
    - Other underdocumented, polysynthetic languages may benefit by applying the *virtuous cycle*

References

Anders Apassingok, Willis Walunga, and Edward Tennant, editors. *Lore of St. Lawrence Island — Echoes of our Eskimo Elders.* BSSD, 1985, 1987, 1989. 3 volumes.

Anders Apassingok, Jessie Uglowook, Lorena Koonooka, and Edward Tennant, editors. *A Reading Series in Yupik and English.* BSSD, 1993, 1994, 1995. 3 volumes.

Linda Womkon Badten, Vera Oovi Kaneshiro, Marie Oovi, and Christopher Koonooka. *St. Lawrence Island / Siberian Yupik Eskimo Dictionary.* ANLC, 2008.

Emily Chen and Lane Schwartz. A morphological analyzer for St. Lawrence Island / Central Siberian Yupik. In *Proceedings of the 11th LREC*, Miyazaki, Japan, May 2018.

Willem J. de Reuse. *Siberian Yupik Eskimo — The Language and Its Contacts with Chukchi.* Studies in Indigenous Languages of the Americas. University of Utah Press, 1994.

Mans Hulden. Foma: a finite-state compiler and library. In *Proceedings of the 12th Conference of the European Chapter of the ACL*, pages 29–32, 2009.

B. Hunt, E. Chen, L. Schwartz, and S. Schreiner. Introducing an electronic dictionary for Central Siberian / St. Lawrence Island yupik. In *Proc. ICLDC*, Feb 2019.

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

Christopher Koonooka. *Ungipaghaghlanga: Let Me Tell You A Story.* Alaska Native Language Center, Fairbanks, Alaska, 2003.

Michael Krauss. St. Lawrence Island Eskimo phonology and orthography. *Linguistics: An International Review*, 13(152):39–72, January 1975.

Kayo Nagai and Della Waghiyi. *St. Lawrence Island Yupik texts with grammatical analysis*, volume A2-006 of *Endangered Languages of the Pacific Rim.* ELPR, 2001.

Daria Morgounova Schwalbe. Sustaining linguistic continuity in the Beringia: Examining language shift and … sustainability. *Anthropologica*, 59(1):28–43, 2017.

L. Schwartz, E. Chen, B. Hunt, and S. Schreiner. Bootstrapping a neural morphological analyzer for SLI Yupik from a finite-state transducer. In *Proc. ComputEL*, Feb 2019.

Lane Schwartz and Emily Chen. Liinnaqumalghiit: A web-based tool for addressing orthographic transparency in St. Lawrence Island Yupik. *LD&C*, 11:275–288, Sept 2017.

Acknowledgements

Portions of this work were funded by NSF Documenting Endangered Languages Grants #BCS 1761680 and 1760977, a University of Illinois Graduate College *Illinois Distinguished Fellowship*, a George Mason University Mathy Junior Faculty Award in the Arts and Humanities, and George Mason University Presidential Scholarship. Special thanks to the Yupik speakers who have shared their language and culture with us.