A Digital Corpus of St. Lawrence Island Yupik
~ for the Yupik Community ~

Lane Schwartz
Department of Linguistics
University of Illinois
lanes@illinois.edu

Emily Chen
Department of Linguistics
University of Illinois
echen41@illinois.edu

Hyunji Hayley Park
Department of Linguistics
University of Illinois
hpark129@illinois.edu

Edward Jahn
ejahn3141@gmail.com

Sylvia L.R. Schreiner
Linguistics Program
Department of English
George Mason University
sschrei2@gmu.edu

INTRODUCTION

- St. Lawrence Island Yupik (Inuit-Yupik, ISO 639-3: *ess*): endangered, polysynthetic language
- Spoken on St. Lawrence Island, Alaska
  Chukotka, Russia
  the Alaskan mainland
  ~1000 speakers total
- Experiencing rapid language shift among younger generations, which have far fewer speakers

- First publicly available digital corpus of written texts in St. Lawrence Island Yupik
- Coordinated with the Native Village of Gambell, the Bering Strait School District, and the Alaska Native Language Center (University of Alaska, Fairbanks)
- Available on GitHub under a Creative Commons Attribution No-Commercial 4.0 International License

GOALS

- Make existing Yupik-language and Yupik pedagogical materials easily and broadly accessible to the community
- Support the development of language technologies (spell-checkers, text-completion software, language learning apps) for use by the community

DIGITIZATION

STEP 1: Scanning

- Scanned texts located at the Gambell School on St. Lawrence Island and at the Alaska Native Language Archive in Fairbanks (600 DPI, TIFF format)

STEP 2: Image Processing

- Deskewed, despeckled, and cropped scans using ScanTailor

STEP 3: Optical Character Recognition

- Performed OCR using ABBYY FineReader

STEP 4: Saving Documents in Accessible Formats

- Saved texts in three formats:
  - Microsoft Word DOCX (for staff at St. Lawrence Island Schools)
  - Searchable PDF/A (for archiving at the Alaska Native Language Archive)
  - UTF-8 Plain Text (for the digital corpus)

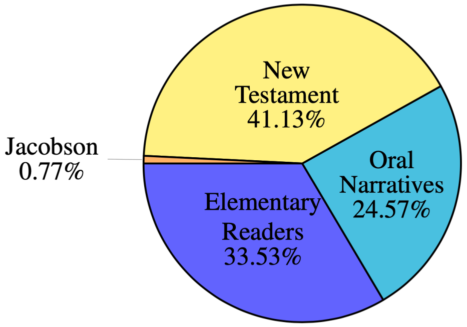

Fig. 1: Distribution of total Yupik sentences per collection, excluding front- & back-matter and English content.

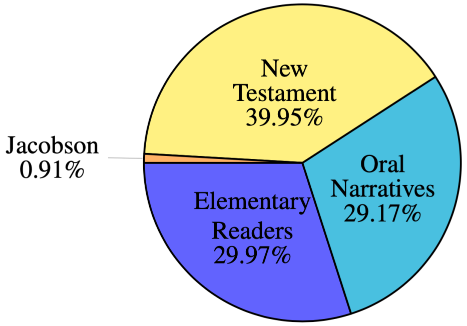

Fig. 2: Distribution of total Yupik word types per collection, excluding front- & back-matter and English content.

THE CORPUS

- Elementary Readers (developed by the Nome Agency Bilingual Education Resource, the Alaska Native Language Center, and the Bilingual Materials Development Center at the Gambell School)
- Oral Narratives (Apassingok et al., 1985, 1987, 1989; Koonooka, 2003; Nagai, 2001; Slwooko, 1977, 1979)
- Practice exercises from Stephen A. Jacobson's 2001 Yupik reference grammar
- New Testament translation and other religious texts, including a hymn book

RELATIONSHIPS

- This work has been undertaken via the collaboration of graduate and undergraduate students, faculty, and community volunteers across two universities
- Our relationships with key individuals in the Yupik language community, including the high school Yupik teacher, several elementary level Yupik teachers, and several elders, have been essential to our development of the corpus and language technologies
- This has allowed corpus materials to already be put to use
- Maintaining these relationships via social media, etc. has made it possible to continue our work during the pandemic

IMPACTS AND OUTCOMES

- Many members of the community have expressed a desire for strengthened Yupik instruction in the school
- This corpus supports that effort by making existing Yupik-language texts available to Yupik educators and community members
- It supports our development of language technology as requested by the community
- It supports research into the structure of Yupik beyond what is currently documented (e.g. Hunt's 2020 corpus study of the word order patterns in quantifier-noun constructions) which may have pedagogical applications in the future
- Lastly, this corpus, in conjunction with our morphological analyzer (Chen et al. 2020), has allowed us to "discover" previously undocumented morphemes and the flexibility of established rules

Selected References

Chen, Emily, Hyunji Hayley Park, and Lane Schwartz. 2020. Improving finite-state morphological analysis for St. Lawrence Island Yupik with paradigm function morphology.

Hunt, Benjamin. 2020. Distributional characteristics of lexical nominal quantifiers in St. Lawrence Island Yupik. George Mason University ms.

Jacobson, Steven A. 2001. *A practical grammar of the St. Lawrence Island/Central Siberian Yupik Eskimo Language*. Fairbanks, AK: Alaska Native Language Center, University of Alaska Fairbanks.

Koonooka, Christopher (Petuwaq). 2005. Yupik language instruction in Gambell (St. Lawrence Island, Alaska). Études/Inuit/Studies, 29(1/2):251–266.

Nagai, Kayo. 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.

Slwooko, Grace. 1977. *Sivuqam Ungipaghaatangi I*. University of Alaska, Anchorage, AK.

Wycliffe. 2018. *Yupik New Testament*. Wycliffe Bible Translators, Saint Lawrence Island, AK.