A Digital Corpus of St. Lawrence Island Yupik

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

Abstract

St. Lawrence Island Yupik (ISO 639-3: *ess*) is an endangered polysynthetic language in the Inuit-Yupik language family indigenous to Alaska and Chukotka. This work presents a step-by-step pipeline for the digitization of written texts, and the first publicly available digital corpus for St. Lawrence Island Yupik, created using that pipeline. This corpus has great potential for future linguistic inquiry and research in NLP. It was also developed for use in Yupik language education and revitalization, with a primary goal of enabling easy access to Yupik texts by educators and by members of the Yupik community. A secondary goal is to support development of language technology such as spell-checkers, text-completion systems, interactive e-books, and language learning apps for use by the Yupik community.

1 Introduction

St. Lawrence Island Yupik (ISO 639-3: *ess*) is an endangered polysynthetic language in the Inuit-Yupik language family (see Figure 1). It is spoken on St. Lawrence Island, Alaska and the Chukotka Peninsula of Russia. This work presents the first publicly available digital corpus of written texts in St. Lawrence Island Yupik, as well as the step-by-step process by which it was created. We refer to this process as our *digitization pipeline*, which can be readily adapted to any other language with any amount of written text.

The public release of the digital corpus has been coordinated with various stakeholders in the St. Lawrence Island community, including the Native Village of Gambell, the Bering Strait School District, the Alaska Native Language Center at the University of Alaska Fairbanks, and Wycliffe Bible Translators.

The digital corpus is now available in plain-text format under the terms of the Creative Commons Attribution No-Commercial 4.0 International License at https://github.com/SaintLawrenceIslandYupik/digital_corpus. Searchable PDF files are being archived at the Alaska Native Language Archive. A mobile-friendly web-accessible version of the corpus will be subsequently developed to allow convenient on- or offline access to the corpus by members of the St. Lawrence Island Yupik community.

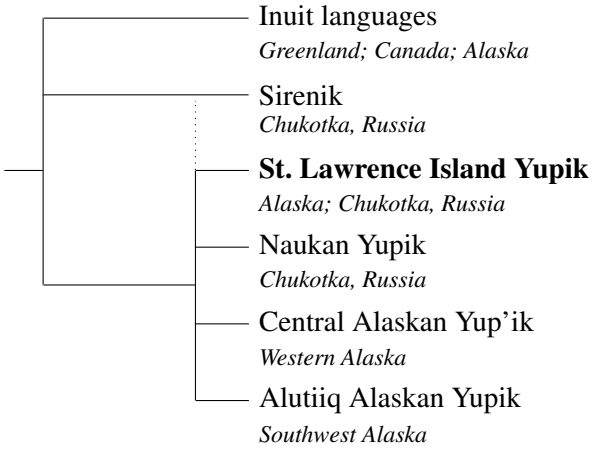

Figure 1: Inuit-Yupik language family (Fortescue et al., 2010; Krauss et al., 2011)

2 Goals for the Corpus

While the vast majority of St. Lawrence Islanders born in or prior to 1980 are fluent L1 Yupik speakers (Krauss, 1980), rapid language shift is underway among younger generations, especially in Russia where language shift is even further advanced (Morgounova, 2007). As a result, many members of the Yupik community have stated a desire for substantially strengthened Yupik instruction in the schools, ideally in the form of a Yupik language immersion program. One obstacle to this, however, is that many Yupik-language texts

Proceedings of the 4th Workshop on the Use of Computational Methods in the Study of Endangered Languages: Vol. 2 Resource Papers and Extended Abstracts, pages 31–40,
Online, March 2–3, 2021.

as well as the pedagogical materials that were developed in the Soviet Union in the early 20th century (Krauss, 1971; Krupnik and Chlenov, 2013) and in Alaska in the late 20th century (Krauss, 1971; Koonooka, 2005) are not easily or broadly accessible. Many materials are also archived at the Alaska Native Language Archive at the University of Alaska Fairbanks and at the Materials Development Center in the Gambell school. Therefore, a primary goal for the development and release of this digital corpus is to strengthen opportunities for Yupik language revitalization and education by enabling easy access to existing Yupik-language texts by educators and by members of the Yupik community. A related secondary goal is to support the development of language technologies such as spell-checkers, text-completion systems, interactive e-books, and language learning apps for use by the community.

3 Digitization Pipeline

We introduce in this section the digitization pipeline used in the creation of the digital corpus. It consists of the following three steps, and can be easily replicated for other languages, since very few aspects of the pipeline were specially tailored to Yupik:

1. scanning
2. image processing
3. optical character recognition

All of the texts that appear in the digital corpus are in UTF-8 plain-text format.

In the United States, Yupik is written using a Latin-derived orthography, while in Russia Yupik is written in a modified Cyrillic orthography. The steps described in this section were applied to Yupik documents created in Alaska written in the Latin-derived Yupik alphabet. A substantial amount of unscanned Cyrillic-orthography Yupik documents were gathered from Soviet libraries and archived at the Alaska Native Language Archive by Krauss (1971); when the global COVID-19 pandemic situation once again allows for safe travel, we plan to scan and process these Cyrillic-orthography Yupik documents using essentially this same pipeline.

3.1 Scanning

During fieldwork visits to Gambell in 2017–2019, we identified and digitized a significant portion of the existing Yupik-language texts. Priority was given to material most likely to be immediately useful in Yupik education efforts and in the development of Yupik language technologies, such as bilingual Yupik-English storybooks.

We gathered all Yupik language materials that could be found in the Gambell school library and Materials Development Center. Most texts were scanned one page at a time using flatbed scanning equipment, while others were scanned using a sheet-fed scanner with an automatic page feeder feature in the Gambell school office. There were a number of texts located at the Alaska Native Language Archive in Fairbanks that were not found in Gambell, and those were scanned on-site in early 2019. Texts were scanned at a resolution of 600 DPI, and whenever possible saved in TIFF image format.

3.2 Image Processing

The raw TIFF image files were processed before optical character recognition was performed. Any images that contained two physical pages were split into two separate files. Next, images were deskewed, despeckled, and cropped. In most cases, these steps were performed using ScanTailor,[1] an open source program designed for such image processing. More recently, we have begun performing these image processing steps directly in ABBYY FineReader,[2] a commercial application that we also use for performing optical character recognition.

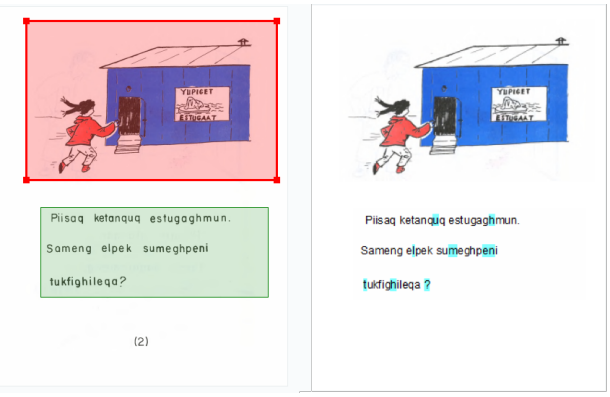

Figure 2: Regions of images and text are identified by ABBYY (left - red and green rectangles, respectively). Low-confidence characters are highlighted during OCR (right - cyan highlights).

[1] https://github.com/scantailor/scantailor

[2] https://pdf.abbyy.com

```
NAGATEK

Ulimakat Nuum Agencym Mumiqhquqhyiqani ,
Nuum Alaskami 99762
Atughqaaluki Title VII-nem Maalghustun
akuzillghestun liinnaqfiganun Bureau of
Indian Affairenun .
```

Figure 3: Sample plain text file of the Yupik front matter from the elementary reader **Nagatek** '*Listen*'.

```
Nagaten .
Nagaqughsigu-u ?

Nakaa .
Sangaawa ?

Esghaqaghhuqun tazigna .

Maaten nagaqughaqa .
Enta aqfaatelta tazingavek .

Hilikaptera .
```

Figure 4: Sample plain text file of the Yupik content from the elementary reader **Nagatek** '*Listen*'.

3.3 Optical Character Recognition

Optical character recognition (OCR) is the process of converting an image into text. As we began the process of scanning Yupik texts in 2016 and 2017, we first attempted to make use of the open source Tesseract[3] OCR software to convert the scanned images into text. While Tesseract models can be trained for new languages, such training requires existing digitized texts. This resulted in a bootstrapping dilemma; without existing Yupik digital texts, we could not train Tesseract models for Yupik.

After poor initial results with Tesseract, we made the decision to switch to ABBYY FineReader (hereafter ABBYY), a state-of-the-art commercial OCR application, for converting our processed image files to plain text. This software was available to us through our respective libraries at the University of Illinois and George Mason University. ABBYY FineReader includes pre-trained OCR models for the broader Inuit-Yupik language family in both Latin and Cyrillic orthographies. Initial work was performed using

[3] https://github.com/tesseract-ocr/tesseract

```
LISTEN

Written and Designed by Myra Poage

Resource Staff/Translators
Raymond Oozevaseuk
Henry Silook
Linda S. Gologerqen

Illustrated by Michael S. Apatiki

A producation of the Nome Agency
Bilingual Education Resource Center,
P.O. Box 1108
Nome, Alaska 99762
for the Title VII Bilingual Education
Program of the Bureau of Indian Affairs

Siberian Yupik
Printed at the GSA Printing Plant
P.O. Box 1612, Juneau, Alaska 99802

May 1975
150 copies
```

Figure 5: Sample plain text file of the English front matter from the elementary reader **Nagatek** '*Listen*'.

```
1.  Listen.
Do you hear it?
2.  No.
What is it?
3.  Look over there.
4.  Now I hear it.
Let's run over there.
5.  Helicopter.
```

Figure 6: Sample plain text file of the English content from the elementary reader **Nagatek** '*Listen*'.

ABBYY version 12, with later work performed using ABBYY version 14. Unlike our early attempts with Tesseract, OCR quality in both versions of ABBYY FineReader was acceptable.

To begin an ABBYY OCR project, all of the TIFF images associated with a document are imported, analyzed, and partitioned into regions that contain text and regions that contain images or illustrations. These regions must be verified, and can be corrected where necessary.

For each text region, we use ABBYY's built-in support for texts written in "Eskimo Latin" to per-

	Yupik			English		
Corpus	**# Sentences**	**# Tokens**	**# Types**	**# Sentences**	**# Tokens**	**# Types**
Total	41,060	268,299	87,102	18,172	202,481	28,619
Elementary Readers	13,402	77,758	25,565	5,643	59,103	8,329
(front- & back-matter)	818	2,364	1,053	962	7,903	2,429
Oral Narratives	9,818	64,696	24,883	10,374	120,194	12,516
(front- & back-matter)	275	1,149	760	886	12,909	4,581
Jacobson Exercises	307	907	772	307	2,372	764
New Testament	16,440	121,425	34,069			

Table 1: Counts of sentences, word tokens, and word types for texts included in the digital corpus. Some elementary readers and oral narratives include front-matter and/or back-matter. Note that the number of sentences (and tokens) in the Yupik and English corpora is not directly comparable - in the Yupik texts lines containing multiple sentences have been split apart and punctuation has been tokenized; in the English texts neither of these steps has yet been performed.

form OCR, and rely on this language setting for all Latin-orthography Yupik documents. We have observed good OCR results even for documents that have deteriorated somewhat over time. ABBYY will nevertheless identify low-confidence characters in the recognized text, and present them to the user for validation. For each section of text that includes one or more low-confidence characters, the TIFF image associated with that section is presented to the user beside a pre-populated text box, in which low-confidence characters are highlighted. The user then confirms or corrects each of these characters. These aspects of image analysis and OCR are shown in Figure 2.

Each fully OCR'd document is then saved in three file formats:

- Microsoft Word DOCX
- searchable PDF/A
- UTF-8 plain text

The Microsoft Word documents will be shared with instructional staff at the St. Lawrence Island schools. Searchable PDF/A files will be archived at the Alaska Native Language Archive, and plain text files are included in our digital corpus.

Lastly, the plain text files are subsequently separated and saved as four individual files. The first file contains any Yupik-language front-matter and back-matter, including title page, table of contents, and appendices (Figure 3). The second file contains the main body of the Yupik text, excluding any front-matter and back-matter (Figure 4). The third file contains the English front-matter and back-matter, if any (Figure 5), and the fourth file contains the English translation of the main body of the text, if any (Figure 6). Furthermore, each sentence of a file appears on its own line, a blank line is used to delimit paragraphs, and punctuation marks are separated out from each line of text. We formatted the files of the digital corpus in this way to ensure that there not only exists a record of the full text, but to also facilitate any NLP work that uses this corpus as a source of data. Separating each text into four individual files enables researchers to easily access the desired data which typically does not include front and back matter. The formatting of each individual file is likewise intended to facilitate text processing.

4 The Digital Corpus of Yupik

To date, we have digitized most of the existing Yupik-language texts using the digitization pipeline introduced herein. We have scanned 90 mostly comb-bound Yupik elementary readers, 7 collections of Yupik oral narratives, the end-of-chapter exercises from the Jacobson (2001) grammar, and 14 collections of Yupik-language hymns and other religious texts. Table 1 summarizes the distribution of sentences, word types, and word tokens across the current digital corpus. We describe each type of text in the following sections.

4.1 Elementary-Level Readers

In the 1970s, a set of elementary-level readers were developed by the Nome Agency Bilingual Education Resource Center at the Bureau of Indian Affairs and by the Alaska Native Language Center at the University of Alaska Fairbanks. In the 1980s, additional readers were developed by

the Bering Strait School District's Bilingual Materials Development Center (MDC) at the Gambell School on St. Lawrence Island. In the early 1990s, a series of five bilingual Yupik-English readers was planned for use by the St. Lawrence Island Schools in grades 4-8 (Apassingok et al., 1993). Only the first three books in the series (Apassingok et al., 1993, 1994, 1995) were actually produced.

To date, 90 of these elementary-level readers have been scanned. Of those, 68 have been fully digitized and are included in the digital corpus, including the bilingual grade 4-6 readers. Processing of the remaining 22 elementary-level readers is ongoing. As seen in Figures 7 and 8, while the elementary-level readers comprise nearly half of the sentences in the digital corpus, they contribute far fewer word types. This is to be expected, given the nature of these texts; since they were originally intended for language learning, one would expect them to frequently repeat words.

4.2 Oral Narratives

In the late 1970s, two books collecting St. Lawrence Island legends were produced by the National Bilingual Materials Development Center at the University of Alaska Anchorage (Slwooko, 1977, 1979). In the 1980s, a set of Yupik oral narratives were recorded on cassette tape, a subset of which were transcribed, translated, and collected into a series of three books constituting the Lore of St. Lawrence Island collection (Apassingok et al., 1985, 1987, 1989). In the late 1990s, a collection of oral narratives were recorded in Savoonga, Alaska as part of fieldwork conducted by Japanese linguist Kayo Nagai. Transcriptions of these narratives, along with interlinear glosses and free translations, were later published in book form (Nagai, 2001). In the early 2000s, a 20th-century collection of Yupik oral narratives from Chukotka was transliterated from Cyrillic into the Latin Yupik orthography and published with English translations (Koonooka, 2003). The oral narratives in these collections include short stories, legends and folktales orally narrated by Yupik elders.

To date, five of these collections of Yupik oral narratives have been fully digitized and are included in the digital corpus, while processing of the remaining two (Slwooko, 1977, 1979) is ongoing. As seen in Figures 7 and 8, the oral narratives contribute approximately one third of the sentences and word types in the digital corpus.

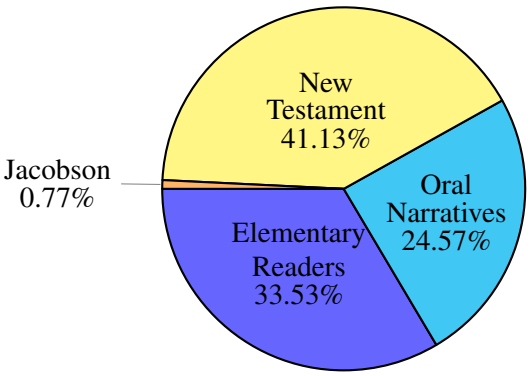

Figure 7: Distribution of total Yupik sentences per collection, excluding front- & back-matter and English content.

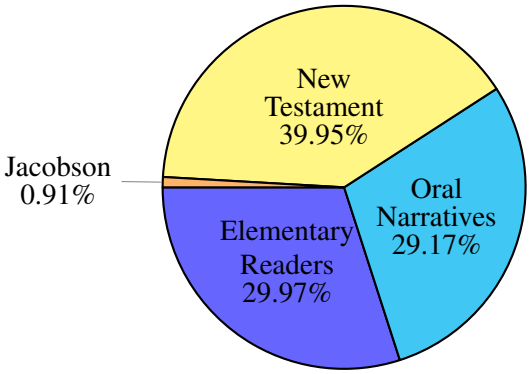

Figure 8: Distribution of total Yupik word types per collection, excluding front- & back-matter and English content.

4.3 Jacobson (2001) End-of-Chapter Exercises

The most thorough source of language documentation for Yupik is the grammar of Jacobson (2001). The grammar is written at the level of an undergraduate college text, and appears to be designed for an audience of L1 Yupik speakers studying their own language at the college level. Chapters 3–17 of the grammar each include end-of-chapter sample Yupik sentences. These sentences are designed for the reader to practice the aspects of Yupik grammar presented in each respective chapter by translating the sentences into English. These end-of-chapter sample sentences have been fully digitized and are included in the digital corpus, though they comprise only a small portion of the corpus as seen in Figures 7 and 8. As part of our Yupik research, we elicited English

reference translations of the Jacobson sample sentences which we also include in the digital corpus.

4.4 Religious Texts

A Yupik translation of the New Testament was published in 2018, completing a nearly 60-year collaborative translation project by Wycliffe Bible Translators and Yupik translators on St. Lawrence Island. There are 14 additional Yupik religious texts (including a collection of hymns) from the Alaska Native Language Archive that we have scanned but not yet fully processed.

5 Corpus's Potential for Linguistic Study

Given Yupik's status as an understudied language, there is no doubt much to be learned linguistically from analyzing this digital corpus. While the corpus has yet to be annotated, preliminary work has yielded several remarkable facts of the language that were not known to us previously, particularly with respect to its morphology.

The morphology of Yupik is perhaps one of the more well-documented aspects of the language. The Yupik lexicon is broadly composed of three types of morphemes: roots, derivational morphemes, and inflectional morphemes. Since Yupik is strictly suffixing with the exception of one prefix, words typically have the following form:

root-derivational morpheme(s)-inflectional morpheme

Most roots are nominal or verbal, such as *alquutagh-* 'spoon' and *qepghagh-* 'to work' respectively. This results in four types of derivational morphemes:

- N→N which suffix to nominal roots and yield nominal stems
- N→V which suffix to nominal roots and yield verbal stems
- V→V which suffix to verbal roots and yield verbal stems
- V→N which suffix to verbal roots and yield nominal stems

Upon suffixation, there are several morphophonological processes that can occur depending on the phonological shape of the root and the morpheme being suffixed. For instance, in (1), suffixing the derivational morpheme *-peragh-* results in the deletion of root-final consonant *-g-* in the root *atkug-*.

(1) **atkuperaq**
atkug-peragh-Ø
parka-makeshift-ABS.sg
'*makeshift parka*' (Badten et al., 2008, p.661)

The Badten et al. (2008) Yupik-English dictionary comprehensively documents all of the morphemes that have been identified to date, while the Jacobson (2001) reference grammar and de Reuse (1994) overview the known ordering constraints (e.g. V→V derivational morphemes may only suffix to verbal roots and stems) and all of the morphophonological processes that can occur upon suffixation. The corpus, however, has demonstrated several exceptions to this documentation.

For example, the derivational morpheme *-pag-* is an augmentive V→V suffix meaning 'to V intensively, excessively'. As such, it is attested to suffix to verbal stems only. One would thus expect it to yield the word seen in (2) but not in (3), where the stem *alquutagh-* 'spoon' is a nominal stem.

(2) **qepghaghpagtuq**
qepghagh-pag-tu-q
work-AUG-IND.INTR-3sg
'*he worked hard*' (Badten et al., 2008, p.659)

(3) **alquutaghpaget**
alquutagh-pag-et
spoon-AUG-ABS.pl
'*heaping tablespoons*' (Nagai, 2001, p.103)

Nevertheless, (3) is a valid, attested form in our digital corpus, and *-pag-* frequently appears suffixed to other nominal stems as well (*sagneghpaget* 'large bowls', *neqekrangllaghpagni* 'great smell of bread', *suupeliighpagni* 'great smell of stew' (Apassingok et al., 1993)). This suggests that perhaps *-pag-* is not only a V→V suffix, but also an N→N suffix, which is not attested in the existing documentation.

A second interpretation is that there exist fewer constraints on morpheme ordering than previously believed, which would permit V→V suffixes to affix to nominal roots. This is also supported by substantial evidence of verbal roots being inflected for nominal inflectional morphemes in our corpus, as

seen in (4).

(4) **yuvghiiq**
yuvghiiq-Ø
examine-ABS.sg
'*look!*' (NABERC, 1975)

In the same way one would not expect the V→V suffix *-pag-* to suffix to a nominal root, one would not expect a verbal root to inflect for nominal inflectional endings. In this way, the digital corpus has opened up a rich area of inquiry.

The digital corpus also speaks to the influence of Yupik's oral tradition. Since many of the texts in our corpus were originally oral narrations, there is considerable speaker variation, which has resulted in transcriptions of morphemes that differ greatly from their attested forms in the Badten et al. (2008) dictionary. These variations are apparent in the morphophonology as well, particularly in regards to allomorphy, as seen in (5).

(5) **maklaguugut**
maklagu-u-gu-t
bearded.seal.intestines-be-IND.INTR-3pl
literal trans. *they are bearded seal's intestines*
(Nagai, 2001, p.191)

(6) **maklagunguut**
maklagu-ngu-u-t
bearded.seal.intestines-be-IND.INTR-3pl
literal trans. *they are bearded seal's intestines*

Whereas the word form in the digital corpus is *maklaguugut*, the word form predicted by the attested morphophonological processes of Yupik is *maklagunguut*, seen in (6). Detailed analysis of word forms such as these would contribute to an increased understanding of Yupik morphophonology, and remedy gaps in the existing documentation.

Our digital corpus thus offers many possibilities for future research in and documentation of Yupik morphology. It would not only facilitate studies on morpheme ordering constraints and morphophonological variation, but would also allow for the potential discovery of novel, previously unattested morphemes. Beyond the level of the word, the corpus will be of use for the study both of morphemes in context and of phenomena at the level of the sentence or the discourse. For example, Yupik boasts a large number of morphemes with meanings related to tense, aspect, mood, and modality. The meanings and uses of these morphemes are not well documented to begin with; in addition, their use often depends on factors outside of the word or sentence they are in. Sentential and discourse-level context is essential for understanding and analyzing many other syntactic and semantic phenomena, as well.

6 Corpus's Potential for NLP Research

Many polysynthetic languages, such as Yupik, are low-resource and under-researched within the field of NLP. The availability of the digital corpus for Yupik now enables researchers to utilize a written dataset that was otherwise inaccessible. For our own purposes the digital corpus has had an immediate impact on two of our projects related to NLP, that is, our ongoing development of a morphological analyzer and a dependency treebank for Yupik.

Given Yupik's rich morphology, the implementation of a morphological analyzer is an essential step in the development of more complex language technologies. While two iterations of a rule-based morphological analyzer have already been implemented (Chen and Schwartz, 2018; Chen et al., 2020), neither achieve full coverage and provide an analysis for all input items. The digital corpus, however, offers a means of understanding the shortcomings of our existing analyzers.

We have already begun a detailed error analysis of the words in the corpus that cannot be analyzed, and are working on identifying the prominent patterns in these errors. For instance, as described in the previous section, our corpus has demonstrated that constraints on morpheme ordering are perhaps more lax than has been initially documented. Knowing this allows us to appropriately modify the analyzer to take this phenomena into account. All of our findings from studying the digital corpus will subsequently be used to improve the existing analyzers.

The digital corpus is also currently being used to create the first Universal Dependencies (UD) (Nivre et al., 2016) treebank for Yupik. UD provides a crosslingual framework for consistent annotation of dependency grammars across different natural languages. However, the framework has not often been utilized for annotating polysynthetic languages like Yupik. By annotating the digital corpus within the UD framework, we hope to contribute to expanding the framework to annotate

other polysynthetic languages. It would further allow us to utilize existing UD tools (e.g. multilingual UD parser) for comparative linguistic research as well as other NLP tasks like syntactic parsing.

A second goal for the UD treebank project is to better understand the syntactic properties of Yupik and to utilize such knowledge for future NLP tasks. In particular, we can use the treebank to create novel sentences in Yupik, thereby augmenting existing textual data. This would greatly assist those NLP tasks that require considerable quantities of data, such as neural language modeling.

In summary, the digital corpus can help us achieve a better understanding of Yupik morphology and syntax, which in turn, would result in the building of more robust computational models. These computational models would then support the development of educational applications for Yupik revitalization, such as spell-checkers, text-completion systems, interactive e-books, and language learning apps.

7 Future Work

To date, we have digitized all of the Yupik-language materials at the Gambell school and a portion of those archived at the Alaska Native Language Archive (ANLA). There are a number of other materials located in the ANLA, however, that have not yet been included in the digital corpus.

After visiting the ANLA in early 2019, we have identified approximately 65 documents indexed under St. Lawrence Island Yupik or Chaplinski Yupik that remain to be scanned. We have also confirmed that there is a substantial amount of Yupik material at the ANLA that has neither been indexed nor scanned, most of which are Soviet-era Yupik texts (primarily in Cyrillic orthography) collected by Michael Krauss during visits to various libraries in the Soviet Union (Krauss, 1971).

Furthermore, the Yupik examples in Shinen (1982), Silook et al. (1983), de Reuse (1994), Shutt et al. (2014) have not been digitized, nor have the examples in Soviet-era Yupik language documentation (Menovshchikov, 1960, 1962, 1967, 1983). The latter are written in Cyrillic orthography with descriptions in Russian. Future work will entail digitizing all of these materials.

A second objective for the digital corpus is text verification. While ABBYY is the state-of-the-art software for OCR work, errors may still have occurred during the OCR process. As such, we plan to have all digitized texts verified by native speakers.

Lastly, we intend to use the digital corpus to eventually build a *parallel* corpus of Yupik texts and their translations. Many of the texts included in the digital corpus have English translations, while many of the Soviet-era works that we plan to include have Russian translations. One challenge, however, is the fact that many of the translations do not have a one-to-one correspondence with Yupik sentences. In such cases, a single Yupik sentence may be translated as more than one English sentence, or vice versa. The intended parallel corpus will map Yupik sentences one-to-one to their translations, which would facilitate various projects and endeavors in NLP.

8 Conclusion

The Yupik language and the corpus of Yupik written texts described herein represent important components of the linguistic and cultural heritage of the St. Lawrence Island Yupik people. While many of the existing Yupik-language texts have already been fully digitized and are present in our digital corpus, there remains much ongoing and future work. As it stands, however, the digital corpus already lends itself to general linguistic inquiry and research related to NLP. We believe it to be a valuable source of data that would greatly contribute to our understanding of the Yupik language, and moreover, to the field of NLP, as computational research on polysynthetic languages is still relatively scarce. Above all, however, is the fact that the digital corpus has broadened the accessibility of Yupik language materials, which is a pivotal step towards establishing a program for Yupik language education and revitalization.

9 Acknowledgments

The Yupik language is a critical part of the cultural heritage of the Yupik people. We offer our deep gratitude to the people of St. Lawrence Island who have trusted us to work with this material. Special thanks to the Yupik speakers whose words are recorded in this corpus. We wish to thank everyone who assisted in scanning, proofreading, and digitizing this material. Thanks to the

board members and staff of the Native Village of Gambell, the City of Gambell, and Sivuqaq, Inc. Thanks to the Bering Strait School District, the faculty and staff of ANLC and ANLA, Dave and Mitzi Shinen of Wycliffe Bible Translators, Steven A. Jacobson, Kayo Nagai, Willem de Reuse, the staff of Gambell Lodge, Iyaaka (Anders Apassingok), Taayqa (Michael James, RIP), Rob Taylor, Petuwaq (Chris Koonooka) and the current and former faculty and staff of Gambell and Savoonga Schools who developed so many wonderful materials over the years and who supported us in this project. This work was supported by NSF Awards 1761680 and 1760977.

Igamsiqanaghhalek!

References

Anders Apassingok, (Iyaaka), Jessie Uglowook, (Ayuqliq), Lorena Koonooka, (Inyiyngaawen), and Edward Tennant, (Tengutkalek), editors. 1993. *Kallagneghet / Drumbeats*. Bering Strait School District, Unalakleet, Alaska.

Anders Apassingok, (Iyaaka), Jessie Uglowook, (Ayuqliq), Lorena Koonooka, (Inyiyngaawen), and Edward Tennant, (Tengutkalek), editors. 1994. *Akiingqwaghneghet / Echoes*. Bering Strait School District, Unalakleet, Alaska.

Anders Apassingok, (Iyaaka), Jessie Uglowook, (Ayuqliq), Lorena Koonooka, (Inyiyngaawen), and Edward Tennant, (Tengutkalek), editors. 1995. *Suluwet / Whisperings*. Bering Strait School District, Unalakleet, Alaska.

Anders Apassingok, (Iyaaka), Willis Walunga, (Kepelgu), and Edward Tennant, (Tengutkalek), 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.

Anders Apassingok, (Iyaaka), Willis Walunga, (Kepelgu), and Edward Tennant, (Tengutkalek), 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.

Anders Apassingok, (Iyaaka), Willis Walunga, (Kepelgu), and Edward Tennant, (Tengutkalek), 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.

Linda Womkon Badten, Vera Oovi Kaneshiro, Marie Oovi, and Christopher Koonooka. 2008. *St. Lawrence Island / Siberian Yupik Eskimo Dictionary*. Alaska Native Language Center, University of Alaska Fairbanks.

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

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

Michael Fortescue, Steven Jacobson, and Lawrence Kaplan. 2010. *Comparative Eskimo Dictionary with Aleut Cognates*, 2nd edition. Alaska Native Language Center, Fairbanks, Alaska.

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

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

Christopher (Petuwaq) Koonooka. 2003. *Ungipaghaghlanga: Let Me Tell You A Story*. Alaska Native Language Center.

Michael Krauss. 1971. Developing a literature in the language of the Eskimos of St. Lawrence Island. Alaska Native Language Archive Identifier SY970K1971c.

Michael Krauss. 1980. Alaska Native languages: Past, present and future. *ANLC Research Papers*, 4.

Michael Krauss, Gary Holton, Jim Kerr, and Colin T. West. 2011. Indigenous peoples and languages of Alaska. Alaska Native Language Archive Identifier G961K2010.

Igor Krupnik and Michael Chlenov. 2013. *Yupik Transitions — Change and Survival at Bering Strait, 1900-1960*. University of Alaska Press, Fairbanks, Alaska.

G. A. Menovshchikov. 1960. *Eskimosskii iazyk*. Gosudarstvennoe uchebno-pedagogicheskoe izdatel'stvo, Leningrad. Pedagogical grammar, similar in scope and level to Jacobson (2001).

G. A. Menovshchikov. 1962. *Grammatika iazyka aziatskikh eskimosov (Grammar of the language of Asian Eskimos)*, volume 1. Izdatel'stvo akademii Nauk (Academy of Sciences of the USSR), Moscow and Leningrad.

G.A. Menovshchikov. 1967. *Grammatika iazyka aziatskikh eskimosov*, volume 2. Izdatel'stvo akademii Nauk, Moscow and Leningrad. Major reference grammar.

G.A. Menovshchikov. 1983. *Slovar' ekimossko-russkiy i russkio-eskimosskiy.* Prosveshchenie., Leningrad. Yupik to Russian and Russian to Yupik School dictionary.

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

Kayo Nagai. 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.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666, Portorož, Slovenia. European Language Resources Association (ELRA).

Nome Agency Bilingual Education Resource Center. 1975. *Aghnaaneq Neghsaq Teghigniqelghii*. GSA Printing Plant, Juneau, AK.

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

David C. Shinen. 1982. Some beginning conversational phrases in St. Lawrence Island Yupik Eskimo. Alaska Native Language Center Identifier SY960S1982.

Lauren Shutt, Dawn Biddison, and Christopher Koonooka. 2014. *Listen & Learn: St. Lawrence Island Yupik Language and Culture Video Lessons*. Arctic Studies Center, Smithsonian Institution, Anchorage, Alaska.

Roger Silook, Adelinda Womkon Badten, Helen Slwooko Carius, Vera Oovi Kaneshiro, Elinor Oozeva, David Shinen, and Grace Slwooko. 1983. *Sivuqam Anglinghhaan Akuzisii / St. Lawrence Island Junior Dictionary.* National Bilingual Materials Development Center, Rural Education Affairs, University of Alaska., Anchorage, Alaska. Yupik-to-English school dictionary.

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

Grace Slwooko. 1979. *Sivuqam Ungipaghaatangi II*. University of Alaska, Anchorage, AK.

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