Community lexical access for an endangered polysynthetic language: An electronic dictionary for St. Lawrence Island Yupik

Benjamin Hunt
George Mason University
bhunt6@gmu.edu

Emily Chen
University of Illinois
at Urbana-Champaign
echen41@illinois.edu

Sylvia L.R. Schreiner
George Mason University
sschrei2@gmu.edu

Lane Schwartz
University of Illinois
at Urbana-Champaign
lanes@illinois.edu

Abstract

In this paper, we introduce a morphologically-aware electronic dictionary for St. Lawrence Island Yupik, an endangered language of the Bering Strait region. Implemented using HTML, Javascript, and CSS, the dictionary is set in an uncluttered interface and permits users to search in Yupik or in English for Yupik root words and Yupik derivational suffixes. For each matching result, our electronic dictionary presents the user with the corresponding entry from the Badten et al. (2008) Yupik-English paper dictionary. Because Yupik is a polysynthetic language, handling of multi-morphemic word forms is critical. If a user searches for an inflected Yupik word form, we perform a morphological analysis and return entries for the root word and for any derivational suffixes present in the word. This electronic dictionary should serve not only as a valuable resource for all students and speakers of Yupik, but also for field linguists working towards documentation and conservation of the language.

1 Introduction

St. Lawrence Island Yupik (hereafter *Yupik*) is an endangered, polysynthetic language of the Bering Strait region, spoken primarily on St. Lawrence Island, Alaska, and the Chukotka Peninsula of Russia.[1] It has undergone a radical language shift in the past few decades, with the youngest generation largely abandoning Yupik in favor of English (Koonooka, 2005) and Russian (Morgounova, 2007), respectively.

There is overt community interest on St. Lawrence Island in Yupik language revitalization, specifically in developing modern technologies to facilitate language-learning. We present one such technology resource: a morphologically-aware web-based version of the Badten et al. (2008) Yupik-English dictionary.[2]

2 Motivation & Prior Work

The electronic dictionary is part of a larger effort to digitize and develop resources for Yupik. A major goal of ours is to build an integrated (mobile-friendly) Yupik language portal to eventually provide St. Lawrence Island community members with an easy mechanism to access digitized Yupik print resources that are integrated with dictionary, morphological analysis, and concordance features.

Over the past two years, we have scanned, cleaned, and OCR'd several volumes of existing Yupik resources, including four anthologies of legends and folk tales (Apassingok et al., 1985, 1987, 1989; Koonooka, 2003) and three elementary readers (Apassingok et al., 1993, 1994, 1995), as well as the Yupik reference grammar of Jacobson (2001). We have also scanned (but not yet cleaned and OCR'd) nearly all of the Yupik language pre-primers, primers, and pedagogical materials present in the school library and Materials Development Center archive in Gambell, Alaska.

In addition to initiating this digital corpus for Yupik, we have also begun to implement a suite of computational systems with a wide range of utilities. To date, we have implemented a Yupik finite-state morphological analyzer (Chen and Schwartz, 2018) and a web utility (Schwartz and Chen, 2017) capable of performing orthotactic spell-checking, transliteration between Yupik's Latin and Cyrillic orthographies and IPA, syllabification, and stress-marking. A morphologically-aware web-based

[1]Special thanks to the Native Village of Gambell and the St. Lawrence Island Yupik speakers who have graciously shared their language and culture with us, the Gambell Schools, and the Alaska Native Language Center. This work was supported by NSF Awards 1761680 and 1760977, by a GMU Presidential Scholarship, and by a University of Illinois Graduate College Illinois Distinguished Fellowship.

[2]Demo at https://youtu.be/quPyL3SXsdk

Proceedings of NAACL-HLT 2019: Demonstrations, pages 122–126
Minneapolis, Minnesota, June 2 - June 7, 2019.

(1) **mangteghaghrugllangllaghyunghitunga**

mangteghagh-	-ghrugllag-	-ngllagh-	-yug-	-nghite-	-tu-	-nga
house-	-big-	-build-	-want.to-	-to.not-	-INTR.IND-	-1SG

'*I didn't want to make a huge house*' (Jacobson, 2001, pg. 43)

(2) **angyasqughhalgunghitukung**

angyagh-	-squghhagh-	-leg-	-ngu-	-nghite-	-tu-	-kung
boat-	-small-	-one.that.has-	-to.be-	-to.not-	-INTR.IND-	-1DU

'*We$_2$ don't, or didn't, have a small boat*' (Jacobson, 2001, pg. 43)

Figure 1: Examples of Yupik words and their component morphemes (shown as interlinear glosses).

Figure 2: The primary search interface of the electronic dictionary.

dictionary represents a significant next step in supporting community language revitalization efforts.

3 Morphologically-aware searchable electronic Yupik dictionary

Yupik is a polysynthetic language with a relatively high average number of morphemes per word (Schwartz et al., 2019). Figure 1 shows examples where a single Yupik word containing multiple morphemes constitutes an entire sentence. Any Yupik electronic dictionary must be sensitive to both derivational and inflectional morphology.

Entries for the electronic dictionary were exported from the original FileMaker Pro database files used by the Alaska Native Language Center to create the Badten et al. (2008) print dictionary. Each entry in the print dictionary includes a Yupik morpheme (either a root or a derivational suffix) in both Latin and Cyrillic orthographies, an English definition, and (for many entries) example sentence(s) in which the morpheme appears and/or other notes about word origin or usage. We augment these entries by specifying the part of speech: each Yupik root is marked as noun, verb, particle, or demonstrative, and each Yupik derivational morpheme is marked for its derivation pattern (that is, as attaching to either a noun root or to a verb root, and as yielding either a noun or a verb). The pronunciation of a Yupik word is predictable from its spelling; we therefore also augment each dictionary entry with its predicted pronunciation in IPA to provide additional utility to linguists working with the language.

Our searchable Yupik electronic dictionary is implemented as a static HTML page and basic CSS style sheet, with dictionary search and morphology functions implemented in Javascript (see Figure 2 above). This provides the user access to all entries from the Badten et al. (2008) print dictionary, including roots and derivational morphemes. Internet access and mobile data coverage on St. Lawrence Island and Chukotka is relatively poor and sometimes unreliable; to support use in these environments the electronic dictionary does not require an internet connection to function.

Users can browse all dictionary entries that begin with a particular letter by selecting that letter from the Yupik alphabet displayed above the search box (Figure 3a), or search for Yupik substrings (Figure 3b), uninflected morphemes (Figure 3d), or fully inflected Yupik words (Figure 3e) including those with multiple morphemes (Figure 3f). The search results present all Yupik dictionary entries (both roots and derivational mor-

(a) After the user selects letter *S* from the search interface, all words beginning with that letter are displayed. The first three results are shown here.

sa (ся) [sɑ] *noun root* - what?; something; relative (kin)
*"sameng piyugsin? 'what do you want?'; sanguzin? 'what are you?'; sa tamaghhaan 'everything'; savut 'our relative', 'our thing(s)'; relative case is saam (rather than *sam); the expanded base sangaa- (underlyingly sangau-) rather than sangu- is used for asking what something (3rd person) is; thus: sangaawa? 'what is it?'; cf. sangwaa"*

saa (сā) [sɑː] *particle* - I don't know; it doesn't register in my mind; never mind
this exclamation is often pronounced with vowel as in English "hat"; = saami"

sa- (ся-) [sɑ] *verb root* - to do what?; to do something
"saa? 'what did he do?'; aatkaghten saat aghvingisafki 'what happened to your clothes when you didn't wash them?; ayveq saa guusavgu? 'what did the walrus do when you shot it?'; saaqat? 'what's going on?, what are they doing?'"

(b) The user can enter a search term either in Yupik or in English. Here, partial results are shown for the incomplete Yupik search term *aghna*. **Yupik words containing that substring are displayed.**

aghnagan (аҕнаган) [ɑʁ.nɑ.ɣɑn] *particle* - hurry up
aghnagan uyuq emta 'hurry up you, knowing how you are'"

aghnagh- (аҕнаҕ-) [ɑʁ.nɑʁ] *verb root* - to wear a dress
aghnaghluni 'wearing a dress'; direct verbalization of aghnaq 'woman'

(c) The user may alternatively search in English. Here, partial results are shown for the search term *family*. **Note that derivational suffixes containing the English search term are returned in addition to roots.**

-nkuk / -nkut (-нкук -нкут) [nkuk.n.kut] *noun-elaborating postbase* - N and partner; N and associate(s); N and family
see %(e)nkuk / %(e)nkut

aalghaq (āлҕақ) [ɑːl.ʁɑq] *noun root* - another family in the same clan; the other of a pair of boats cooperating in a hunt; hunting partner; second wife

(d) Yupik searches may also return entries corresponding to Yupik derivational suffixes. Here, partial results are shown for the search term *nkut*.

-nkuk / -nkut (-нкук -нкут) [nkuk.n.kut] *noun-elaborating postbase* - N and partner; N and associate(s); N and family
see %(e)nkuk / %(e)nkut

kinkut (кинкут) [kin.kut] *noun root* - who? (plural)
look under kina

(e) The user may search for fully inflected Yupik words. Here, results are shown for the search term *nagatunga* 'I listened.'

nagate- (нагаты-) [nɑ.ɣɑ.tə] *verb root* - to listen
nagatuq 'he is listened'; nagataa 'he listened to her' /

(f) Preliminary support is included for multi-morphemic Yupik word searches. Here, partial results for the Yupik word *mangteghaghrugllangllaghyunghitunga* 'I didn't want to make a huge house' **are shown.**

mangteghagh- *verb root* - to make a house for; to reassemble a house (in fall after its parts have been aired out)
direct verbalization of mangteghaq 'house'

–ghrugllak *noun-elaborating postbase* - big N; large N
mangteghaghrugllak 'big house', mangteghaghrugllaget 'very big houses' (from mangteghaq 'house')

(g) Preliminary support is included for Yupik searches using the Cyrillic orthography. Here, partial results for the Yupik word *qikmiq* **are shown. The search was performed in Cyrillic.**

qikmighaq (қикмиҕақ) [qik.m̥i.ʁɑq] *noun root* - puppy

qikmiq (қикмиқ) [qik.m̥iq] *noun root* - dog

Figure 3: Dictionary results for various types of searches.

phemes) that completely or partially match the Yupik search string. Users may also search in English. In that case, search results return any dictionary entries that contain the search term as a substring in the English definition (Figure 3c).

In order to facilitate use by Yupik speakers in Chukotka as well as Alaska, preliminary support is included for searches where the Yupik search term is input in the Yupik Cyrillic orthography. Figure 3g depicts results where the search term was written in Cyrillic. Currently, searches performed in Cyrillic return English entries only, as entries pull from the English-language Badten et al. (2008) dictionary. This may be useful for a user who does not speak Russian but who has come across a Yupik word written in Cyrillic and wishes to find out its meaning. However, it does not address the needs of Yupik speakers who also speak Russian, but not English. We intend future iterations to integrate entries from Russian-language dictionaries of Yupik, such that a search performed in Cyrillic will return Yupik or Russian entries, depending on the search term, just as a search performed in the Latin orthography returns Yupik or English entries.

4 Community & Research Impacts

The electronic dictionary with its existing functionalities has the ability to make a significant impact on the Yupik language community as well as researchers working on the language. The current version of the electronic dictionary includes preliminary support for multi-morphemic searches using the integrated morphological analyzer, and preliminary support for searches performed in Cyrillic. More robust implementations of these features are ongoing.

We anticipate that the electronic dictionary will greatly facilitate access to knowledge that was otherwise difficult to obtain, and make it readily available to all community members. Versions of the print dictionary have been available through the University of Alaska Press in various editions since 1983, with revisions in 1987 and 2008. However, the print edition is bulky and relatively expensive; while the school libraries in Gambell and Savoonga have copies, most community members (including some members of the Yupik Bible translation project) do not.

While the dictionary should impact all community members, regardless of age, we expect it to most positively shape the language experience of the younger generations, promoting language use among youth who may be unlikely (for social reasons) to ask elders about word-forms. Moreover, one of the most promising features of the dictionary with respect to language learning is the integration of the `foma` finite-state analyzer, which allows fully inflected word forms with multiple morphemes to be searched in order to either define parts of the word or to reconstruct its full meaning. This should be especially valuable for students who have not yet mastered the polysynthetic aspects of the Yupik language. For example, students could be allowed to use the dictionary in the classroom to help them read through Yupik texts that contain vocabulary that is at a higher level than they might otherwise be able to handle. Students would not need to be able to parse an unfamiliar word to be able to look up the meaning of the root.

The electronic dictionary is practical in much the same ways for linguists and researchers, allowing them to swiftly search for word forms and definitions via a resource that is significantly more portable than a two-volume paper dictionary. The integration of the `foma` finite-state analyzer is of particular note, however, since it can be used in the construction of morphological interlinear glosses (see Figure 1), which are critical for the processing and sharing of linguistic data. The electronic dictionary supplemented with the morphological analyzer greatly expedites this process, which must otherwise be done by hand.

We plan to conduct live user field testing of the electronic dictionary in the Gambell School during spring and summer 2019. User feedback will inform user-interface redesign decisions and will provide valuable feedback regarding which features are most valued by Yupik community members. We also plan to embed the dictionary in native mobile apps for Android and iOS, and to conduct field testing of those user interfaces as well.

While this electronic dictionary provides direct support for the language revitalization efforts of the Yupik community specifically, we hope that it might serve as a blueprint for similar tools for other endangered languages, particularly those of a polysynthetic nature. Such an analyzer-linked dictionary may be of use both to the language communities themselves, and to researchers working with the communities to reinforce their efforts.

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.

Adelinda W. (Aghnaghaghpik) Badten, Vera Oovi Kaneshiro, Marie Oovi, and Steven A. Jacobson, editors. 1983. *A Dictionary of the St. Lawrence Island / Siberian Yupik Eskimo Language*, 1st edition. Alaska Native Language Center, Fairbanks, Alaska. Alaska Native Language Archive Identifier SY975J1983b.

Adelinda W. (Aghnaghaghpik) Badten, Vera Oovi Kaneshiro, Marie Oovi, and Steven A. Jacobson, editors. 1987. *A Dictionary of the St. Lawrence Island / Siberian Yupik Eskimo Language*, 2nd edition. Alaska Native Language Center, Fairbanks, Alaska.

Linda Womkon Badten, (Aghnaghaghpik), Vera Oovi Kaneshiro, (Uqiitlek), Marie Oovi, (Uvegtu), and Christopher Koonooka, (Petuwaq). 2008. *St. Lawrence Island / Siberian Yupik Eskimo Dictionary*. Alaska Native Language Center, University of Alaska Fairbanks. Alaska Native Language Archive Identifier SY975J2008.

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 (LREC'18)*, Miyazaki, Japan.

Steven A. Jacobson. 2001. *A Practical Grammar of the St. Lawrence Island/Siberian Yupik Eskimo Language*, 2nd edition. Alaska Native Language Center, University of Alaska Fairbanks, Fairbanks, Alaska. Alaska Native Language Archive Identifier SY975J2001.

Christopher Koonooka, (Petuwaq). 2003. *Ungipaghaghlanga — Quutmiit Yupigita Ungipaghaatangit / Let Me Tell a Story — Legends of the Siberian Eskimos*. Alaska Native Language Center, University of Alaska Fairbanks, Fairbanks, Alaska. Transliterated and translated from the Chukotka collection of G.A. Menovshchikov. Stories told by Ayveghhaq, Tagikaq, Asuya, Alghalek, Nanughhaq, and Wiri. Alaska Native Language Archive Identifier SY003K2003.

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

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

Lane Schwartz and Emily Chen. 2017. Liinnaqumalghiit: A web-based tool for addressing orthographic transparency in St. Lawrence Island/Central Siberian Yupik. *Language Documentation and Conservation*, 11:275–288.

Lane Schwartz, Sylvia L.R. Schreiner, and Emily Chen. 2019. Community-focused language documentation in support of language education and revitalization for St. Lawrence Island Yupik. *Études/Inuit/Studies*. Forthcoming.