LANGUAGE DOCUMENTATION & CONSERVATION

Vol. 11 (2017), pp. 275–288
http://nflrc.hawaii.edu/ldc
http://hdl.handle.net/10125/24736
Revised Version Received: 28 Aug 2017

Liinnaqumalghiit: A web-based tool for addressing orthographic transparency in St. Lawrence Island/Central Siberian Yupik

Lane Schwartz
University of Illinois

Emily Chen
University of Illinois

We present an initial web-based tool for St. Lawrence Island/Central Siberian Yupik, an endangered language of Alaska and Russia. This work is supported by the local language community on St. Lawrence Island, and includes an orthographic utility to convert from standard Latin orthography into a fully transparent representation, a preliminary spell checker, a Latin-to-Cyrillic transliteration tool, and a preliminary Cyrillic-to-Latin transliteration tool. Also included is a utility to convert from standard Latin orthography into both IPA and Americanist phonetic notation. Our utility is also capable of explicitly marking syllable boundaries and stress in the standard Latin orthography using the conventions of Jacobson (2001), as well as in Cyrillic and in standard IPA notation. These tools are designed to facilitate the digitization of existing Yupik resources, facilitate additional linguistic field work, and most importantly, bolster efforts by the local Yupik communities in the U.S. and in Russia to promote Yupik usage and literacy, especially among Yupik youth.

1. Introduction St. Lawrence Island/Central Siberian Yupik (ISO 639-3: ess), also known as Chaplinski in the Russian literature, is an endangered language of the Bering Strait region indigenous to St. Lawrence Island in far western Alaska and the Chukchi Peninsula of far eastern Russia (Figure 1). The endonym Akuzipik is also sometimes used locally.

This language is the westernmost variety in the Inuit-Yupik language family,[1] which ranges across the Arctic coast of North America from Greenland, through northern Canada, northern and western Alaska, to the Chukchi Peninsula in far eastern Russia. The Yupik branch of this family encompasses St. Lawrence Island / Central Siberian Yupik, Naukan, Central Alaskan Yup'ik, and Sugpiaq (Jacobson 2001: see Figure 2). A fifth (now extinct) language, Sirenik, is thought to belong either to Yupik or to constitute an independent third branch between Yupik and Inuit (Vakhtin

[1] This language family has historically been known as Eskimo. In recent years there has been a trend, especially in Canada, away from the exonym Eskimo, driven in large part by an alleged (and disputed) derogatory etymology (Mailhot 1978).

E-ISSN 1934-5275

& Golovko 1987; Vakhtin 1998). Despite the geographic and political separation, the differences between the varieties of Yupik on St. Lawrence Island and Chukotka are relatively minor (Krauss 1975).

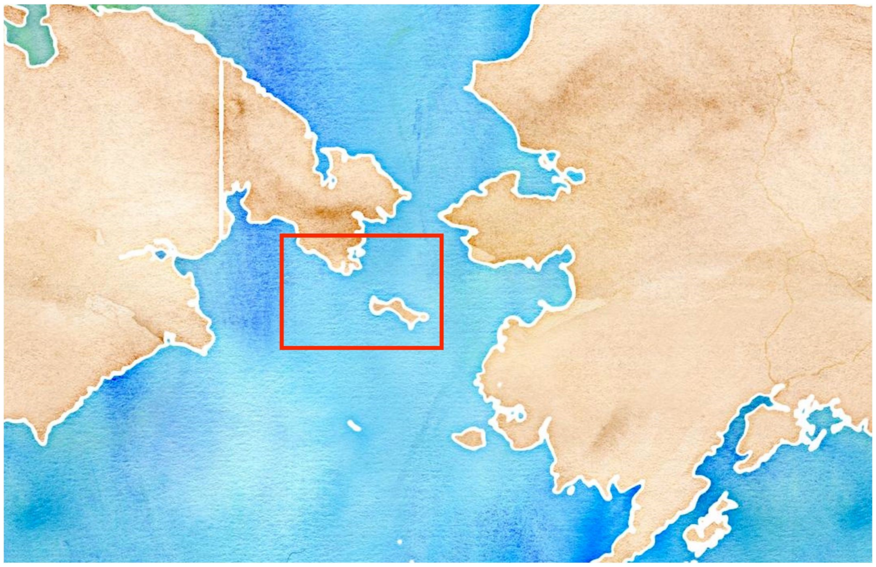


Figure 1. Bering Strait region;[2] rectangle highlights St. Lawrence Island and eastern Chukotka, where Yupik is spoken (Krauss et al. 2010).

While a small number of computational resources have been developed for some languages in the Inuit branch of this language family, to our knowledge, no computational tools have previously been developed for St. Lawrence Island Yupik. We present a web-based tool[3] that includes a preliminary spell checker, a utility to convert from the standard Latin orthography into a fully transparent orthographic representation, and a Latin-to-Cyrillic transliterator. In addition, our work also includes a utility to convert from the standard Latin orthography into both IPA and Americanist phonetic notation. Lastly, we include a utility that explicitly marks syllable boundaries and stress in the standard Latin orthography using the conventions of Jacobson (2001), as well as in Cyrillic and in standard IPA notation. Our implementation has no server-side dependencies, thus enabling classroom use in environments with limited, unreliable, or no internet access.

We call our collection of tools **Liinnaqumalghiit**, meaning "those who are willing to learn." These tools are designed to facilitate the digitization of existing Yupik resources, facilitate additional linguistic field work, and most importantly, to bolster efforts by the local Yupik communities in the U.S. and in Russia to promote Yupik usage and literacy, especially among Yupik youth.

[2] Map generated by http://worldmap.harvard.edu/maps/new using data by OpenStreetMap and tiles by Stamen Design and made available under CC BY SA 3.0 and CC BY 3.0 licenses, respectively.

[3] http://computational.linguistics.illinois.edu/yupik/

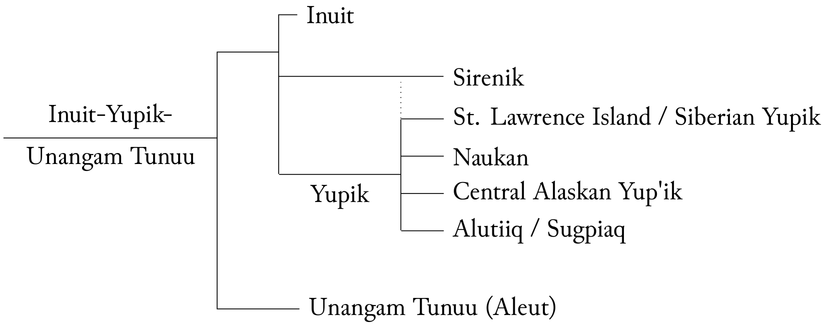


Figure 2. Inuit-Yupik-Unangam Tunuu language family

2. Language status Over the past century, the language has encountered widely varying levels of support from the political and educational systems in Russia and Alaska. In keeping with early Soviet policy that supported its minorities, Central Siberian Yupik was selected as the language of choice for all Soviet Eskimo national literature, and its speakers in Chukotka were encouraged to maintain their ethnic identity and traditions (Krauss 1975). A shift in this policy, however, coupled with increased use of Russian, has resulted in a Yupik-speaking population numbering no more than 300 in Chukotka, with the youngest generation possessing little to no knowledge of Yupik at all (Vakhtin 1998; Morgounova 2007).

In 1972, Yupik was reintroduced into the school curriculum on St. Lawrence Island (Krauss 1975), heralding the publication of language materials including multiple spiral-bound booklet readers and pre-primers prepared by the Bering Strait School District Bilingual Program and the Nome Agency Bilingual Education Resource Center (Tennant 1985). As recently as the 1980s, the vast majority of children on St. Lawrence Island learned Yupik at home from their parents. Yet even with these greater efforts to cultivate and stabilize the language on the Alaskan side, Yupik has had relatively low levels of formal support over the past decade, with substantially increased movement away from Yupik in favor of English by the youngest generation (Koonooka 2005; Schwalbe 2015).

A glance at the existing linguistic corpora of Yupik yields a multitude of ethnographies, translated collections of folk songs and stories (Apassingok & Tennant 1987; Apassingok et al. 1985; 1987; 1989; 1993; 1994; 1995; Koonooka 2003; Shutt et al. 2014), and linguistic fieldwork papers, with extensive contributions from researchers in the former Soviet Union (Krupnik 1983; 1994). Despite the fact that most Yupik texts produced in or after the 1980s were born digital, nearly all Yupik resources are available only in print form, with many of the electronic originals lost or on difficult-to-access aging media/legacy formats.

3. Yupik phonology and orthography The sound inventory of Yupik contains 31 consonants and four vowels, /s/, /i/, /a/, and /u/, the latter three of which can be lengthened for a total of seven unique vocalic phonemes. Yupik allows CV and CVC syllables (where V may be a long vowel), as well as word-initial V and VC syllables.

Yupik therefore does not permit consonant clusters except at syllable boundaries, and moreover unlike all of its sister languages, it has neither diphthongs nor geminate consonants (Jacobson 2001).

On St. Lawrence Island, an attempt at a Latin-based Yupik orthography was devised as early as 1910, although the standard system in use today was not developed until 1971 (Krauss 1975). A parallel Cyrillic-based Yupik orthography was developed in Russia. Both orthographies, along with their IPA representations, are presented in Tables 1 and 2.

3.1 Orthographic undoubling The Yupik consonant inventory includes 8 pairs of continuants that differ only in voicing and 4 pairs of nasal consonants that differ only in voicing; of these, the voicing distinction is systematically marked in the Latin orthography through graphemic *doubling* in 5 of the 8 continuant pairs (**l** and **ll**, **r** and **rr**, **g** and **gg**, **gh** and **ghh**, **ghw** and **ghhw**) and in all of the nasal pairs (**m** and **mm**, **n** and **nn**, **ng** and **ngng**, **ngw** and **ngngw**). For example, the voiced and voiceless velar fricatives are written as **g** and **gg**, respectively. Krauss (1975:66) notes that the standard Yupik orthography used today, which includes these multigraphemic letters, was explicitly designed to eliminate "all diacritics and non-standard symbols," which Krauss claims were "major disadvantages of the previous system."

Yupik has rich morphology, which, combined with this multigraphemic orthography, commonly results in rather long words; for example, in Apassingok et al. (1985:22) we observe the word **iknaqngwaaghuyaghpetut**. In general, Yupik phonology requires that adjacent consonants in a word agree in voicing; thus, consonants in a cluster may either both be voiced or unvoiced (Jacobson 2001).[4] Yupik orthographic conventions take advantage of the phonological rules regarding consonant cluster voicing to shorten the spelling of words containing clusters of unvoiced consonants; in a consonant cluster, when a graphemically doubled voiceless consonant (such as **ngngw**) co-occurs with another voiceless consonant (for example **q**) and the phonology dictates that the entire cluster must be voiceless, the graphemically doubled consonant is replaced in the written form with its undoubled counterpart (such as **ngw**). In our example word, orthographic undoubling occurs in three distinct places:

- the consonant cluster /kn/ is written as **kn** rather than **knn**
- the consonant cluster /qŋ̊ʷ/ is written as **qngw** rather than **qngngw**
- the consonant cluster /χp/ is written as **ghp** rather than **ghhp**

Without this orthographic convention, the word /ikn̥ɑqŋ̊ʷɑːʁujɑχpətut/ would be spelled **iknnaqngngwaaghuyaghhpetut**; because of the orthographic convention, the actual spelling is instead **iknaqngwaaghuyaghpetut**.

[4] Unvoiced consonants may follow voiced nasals.

	Labial	Alveolar		Palatal		Retroflex	Velar	Velar (rounded)	Uvular	Uvular (rounded)	Glottal	
Unvoiced Stops	**p**	**t**					**k**	**kw**	**q**	**qw**		Latin
	p	t					k	k^w	q	q^w		IPA
	п	**т**					**к**	**кӱ**	**ӄ**	**ӄӱ**		Cyrillic
Voiced Continuants	**v**	**l**	**z**		**y**	**r**	**g**	**w**	**gh**	**ghw**		Latin
	v	l	z		j	ɻ	ɣ	$ɣ^w$	ʁ	$ʁ^w$		IPA
	в	**л**	**з**		**й**	**р**	**г**	**(г)ӱ**	**ӷ**	**ӷӱ**		Cyrillic
Unvoiced Continuants	**f**	**ll**		**s**		**rr**	**gg**	**wh**	**ghh**	**ghhw**	**h**	Latin
	f	ɬ		s		ʂ	x	x^w	χ	$χ^w$	h	IPA
	ф	**ль**		**с**		**ш**	**х**	**хӱ**	**ӽ**	**ӽӱ**	**г**	Cyrillic
Voiced Nasals	**m**	**n**					**ng**	**ngw**				Latin
	m	n					ŋ	$ŋ^w$				IPA
	м	**н**					**ң**	**ңӱ**				Cyrillic
Unvoiced Nasals	**mm**	**nn**					**ngng**	**ngngw**				Latin
	m̥	n̥					ŋ̊	$ŋ̊^w$				IPA
	мь	**нь**					**ңь**	**ңьӱ**				Cyrillic

Table 1. Yupik consonant chart, adapted from Krauss (1977), Badten et al. (2008), and Jacobson (2001) in St. Lawrence Island (Latin) orthography, International Phonetic Alphabet (IPA), and Chukotkan (Cyrillic) orthography. Some authors (e.g., Nagai 2001) use an Americanist phonetic notation, which we omit from this table, but include in our conversion utilities.

Close Vowels	**i** i **и**		**u** u **у**	Latin IPA Cyrillic
Mid Vowel		**e** ə **ы**		Latin IPA Cyrillic
Open Vowel		**a** ɑ **а**		Latin IPA Cyrillic

Table 2. Yupik vowel chart Krauss (1975), Badten et al. 2008, Jacobson (2001) in St. Lawrence Island (Latin) orthography, International Phonetic Alphabet (IPA), and Chukotkan (Cyrillic) orthography.

Figure 3. **Function 1** Undouble a consonant. This function is used in Undoubling Rules 1–3 of Jacobson (2001).

```
function UNDOUBLE(c)
                                          ▷ Continuants with undoubled variant
    if c == ll then return l
    else if c == rr then return r
    else if c == gg then return g
    else if c == ghh then return gh
    else if c == ghhw then return ghw
                                          ▷ All nasals can be undoubled
    else if c == mm then return m
    else if c == nn then return n
    else if c == ngng then return ng
    else if c == ngngw then return ngw
    else                                  ▷ No undoubled variant
        return c
    end if
end function
```

In this example, this orthographic undoubling convention takes advantage of the fact that fluent Yupik speakers know that voiced /ŋ^{w}/ cannot follow voiceless /q/, and so **qngw** must be pronounced /qŋ̊^{w}/, even though **ngw** is normally pronounced /ŋ^{w}/. This intuition is similar to the way that native English speakers know that the **-s** at the end of **cats** is voiceless /s/ while the **-s** at the end of **dogs** is voiced /z/ (Jacobson 2001).

Table 3. Jacobson (2001) Undoubling Rule 1 Undouble a continuant when it precedes a voiceless consonant where doubling is not used to show voicelessness.

```
Let [:voiceless:] =     [ p | t | k | kw | q | kw | f | s | wh ]
Let c =                 [ ll | rr | gg | ghh | ghhw ]
c →                     undouble(c) / _ [:voiceless:]
```

Table 4. **Jacobson** (2001) **Undoubling Rule 2** Undouble a nasal when it follows a voiceless consonant where doubling is not used to show voicelessness.

```
1:  Let [:voiceless:] =     [ p | t | k | kw | q | kw | f | s | wh ]
2:  Let c =                 [ mm | nn | ngng | ngngw ]
3:  c →                     undouble(c) / [:voiceless:] _
```

Table 5. **Jacobson** (2001) **Undoubling Rule 3** Undouble a continuant or nasal when it precedes ll or when it follows a voiceless consonant where doubling is used to show voicelessness.

```
1:  Let [:voiceless:] =     [rr | gg | ghh | ghhw | mm | nn | ngng | ngngw ]
2:  Let c =                 [ ll |rr | gg | ghh | ghhw | mm | nn | ngng | ngng w ]
3:  c →                     undouble(c) / _ ll
4:  c →                     undouble(c) / [:voiceless:] _
```

Table 6. Orthographic undoubling. Consonants made transparent in the second line are marked in **bold**.

ik**n**aq**ngw**aaghuya**gh**petut	Latin
ik**nn**aq**ngngw**aaghuya**ghh**petut	Transparent
ik.ʼn̥ɑq.ʼŋ̊ʷɑː.ʁu.ʼjɑχ.pə.tut	IPA

4. Computational tools for Yupik

There is a stated desire by the St. Lawrence Island Yupik community to strengthen and revitalize the use of Yupik (Koonooka 2005). In consultation with the local tribal government and local Yupik instructors on St. Lawrence Island, we have begun a project to develop a set of computational tools for Yupik specifically designed to assist in those goals by facilitating the digitization of existing Yupik printed resources, facilitating additional linguistic field work, and bolstering efforts by the local Yupik communities to promote Yupik usage and literacy.

Our long-term goal is the digitization of existing Yupik corpora within a searchable digital framework connected to an electronic dictionary and morphosyntactic analyzer, with all texts accessible in both Latin and Cyrillic orthographies, for pedagogical use in local Yupik instructional settings and for use in further documentation of the Yupik language.

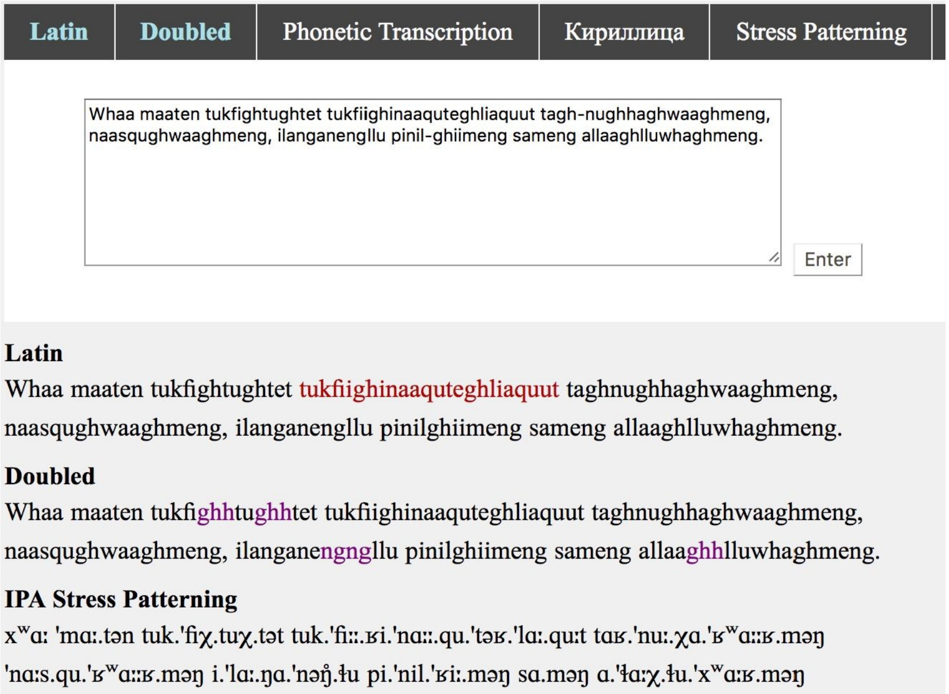
Our short-term goal is the development and release of those tools most likely to be of immediate use by Yupik language instructors on St. Lawrence Island and their students. An open source web implementation of our tools can be accessed online or downloaded for offline use. In this section we present the tools that we have implemented to date.

4.1 Increased orthographic transparency

The first issue we address is the problem posed by orthographic undoubling (§3.1) to students with limited Yupik proficiency.

In Alaska, the sounds of the Yupik sound system are represented in a mostly transparent Latin orthography. The Yupik Latin orthography, along with IPA representations and Cyrillic counterparts, is presented in Tables 1 (consonants) and 2 (vowels). While orthographic undoubling was viewed as intuitive to many fluent speakers when the current Latin orthography was developed (Krauss 1975), the system is not necessarily so to those today learning Yupik at school rather than at home. The first tool we present is designed to alleviate this non-transparency in orthography.

Our technique converts Yupik words from the standard Latin orthographic forms into a transparently pronounceable form, fully resolving any orthographic opacity. This is done by transforming orthographically undoubled graphemes into their (orthographically transparent) doubled counterpart. We begin with a greedy longest-match orthographic tokenizer that reads (in reverse) each Yupik word in Latin orthography and returns an array of (multi-)graphemic characters that are used in the implementation of the other functions we present.



Whaa maaten tukfightughtet tukfiighinaaquteghliaquut tagh-nughhaghwaaghmeng, naasqughwaaghmeng, ilanganengllu pinil-ghiimeng sameng allaaghlluwhaghmeng.

Latin

Whaa maaten tukfightughtet tukfiighinaaquteghliaquut taghnughhaghwaaghmeng, naasqughwaaghmeng, ilanganengllu pinilghiimeng sameng allaaghlluwhaghmeng.

Doubled

Whaa maaten tukfighhtughhtet tukfiighinaaquteghliaquut taghnughhaghwaaghmeng, naasqughwaaghmeng, ilanganengngllu pinilghiimeng sameng allaaghhlluwhaghmeng.

IPA Stress Patterning

xʷɑː 'mɑː.tən tuk.'fiχ.tuχ.tət tuk.'fiː:.ʁi.'nɑː:.qu.'təʁ.'lɑː.quːt tɑʁ.'nuː.χɑ.'ʁʷɑː:ʁ.məŋ 'nɑːs.qu.'ʁʷɑː:ʁ.məŋ i.'lɑː.ŋɑ.'nəŋ̊.ɬu pi.'nil.'ʁiː.məŋ sɑ.məŋ ɑ.'ɬɑːχ.ɬu.'xʷɑːʁ.məŋ

Figure 4. Sample output from our website where the *Latin*, *Doubled*, and *IPA Stress Patterning* options have been selected. Misspelled words are in red, undone undoubling in purple.

Undoubling Rules 1–3 (Krauss 1975; Jacobson 2001) on pages 280–281 describe the process of orthographic undoubling. We iterate over the word's graphemes, identifying those environments where one of Jacobson's Undoubling Rules has taken place, undoing any undoubling by applying our own Doubling Rules 1–3 (page 283). The newly doubled grapheme is highlighted in purple in the web UI to visually accentuate the orthographic revision (see Figure 4). We thus present the user with a fully trans-

Figure 5. **Function 2** Double a previously undoubled consonant.

```
function DOUBLE(c)
                                        ▷ Continuants with doubled variant
    if c == l then return ll
    else if c == r then return rr
    else if c == g then return gg
    else if c == gh then return ghh
    else if c == ghw then return ghhw
                                        ▷ All nasals can be doubled
    else if c == m then return mm
    else if c == n then return nn
    else if c == ng then return ngng
    else if c == ngw then return ngngw
    else                                ▷ No doubled variant
        return c
    end if
end function
```

Table 7. **Doubling Rule 1** Double a previously undoubled continuant when it precedes a voiceless consonant where doubling is not used to show voicelessness.

```
Let [:voiceless:] =    [ p | t | k | kw | q | kw | f | s | wh ]
Let c =                [ l | r | g | gh | ghw ]
c →                    double(c) / _ [:voiceless:]
```

Table 8. **Doubling Rule 2** Double a previously undoubled nasal when it follows a voiceless consonant where doubling is not used to show voicelessness.

```
Let [:voiceless:] =    [ p | t | k | kw | q | kw | f | s | wh]
Let c =                [ m | n | ng | ngw ]
c →                    double(c) / [:voiceless:] _
```

Table 9. **Doubling Rule 3** Double a previously undoubled continuant or nasal when it precedes ll or when it follows a voiceless consonant where doubling is used to show voicelessness.

```
Let [:voiceless:] =    [ rr | gg | ghh | ghhw | mm | nn | ngng | ngngw]
Let c =                [ l | r | g | gh | ghw | m | n | ng | ng w]
c→                     double(c) / _ ll
c→                     double(c) / [:voiceless:] _
```

parent orthographic realization of their input text, replacing all relevant undoubled graphemes with the appropriate doubled variant.

4.2 Transliteration While our computational tools are predominantly designed to assist revitalization and pedagogical efforts on St. Lawrence Island, we hope that they may be of use to the Yupik community in Chukotka as well. In light of this, we have incorporated a Latin-to-Cyrillic transliterator that transliterates each Latin grapheme into its Cyrillic counterpart. While the Latin orthography employs a series of undoubling rules to simplify the writing system, the Cyrillic orthography implements a differing set of orthographic conventions (Jacobson 2001). We implement a rule-based Latin-to-Cyrillic transliteration tool following these conventions.

We have also implemented a preliminary Cyrillic-to-Latin utility. In principle this task is hampered by the fact that Cyrillic letter *г* is used to represent both /ɣ/ and /h/. However, as Krauss (1975) points out, /h/ is a rare phoneme in Yupik; the Badten et al. (2008) dictionary lists 25 bases[5] and zero postbases that include /h/. When transliterating from Cyrillic into Latin, we therefore check for the presence of these bases, and in all other cases transliterate *г* as g.[6]

4.3 Phonetic transcription, stress, and syllabification To further assist field linguists and other scholars of Yupik, we have also included functions to transcribe input text into the standard International Phonetic Alphabet (IPA) and/or the Americanist phonetic notations utilized by some Yupik scholars (Krauss 1975; Nagai 2001). These functions accept tokenized text in the Latin orthography, and transliterate each individual grapheme into its IPA or Americanist phonetic equivalent.

Another feature that may be of use to field linguists, and especially students studying Yupik, is a utility that explicitly marks syllable boundaries and stress patterns according to the rules that govern the relationship between syllables and stress in Yupik. Our syllabifier takes a tokenized Yupik word and marks the syllable boundaries within that word, passing this result to a second function that marks stress according to the following three rules from Jacobson (2001:6–8):

- Alternating vowels are stressed, beginning with the second vowel in a word, although the final syllable does not receive stress.
- If the first vowel in a long vowel is stressed (**aa, ii**, and **uu**), the stress is advanced to the second vowel and the alternating pattern continues from this second vowel, e.g. **ang/yul/kum/tung/llu**.
- In a process called *rhythmic lengthening*, the full vowels, **i, a**, and **u** are lengthened if they are stressed and belong to an open syllable. If this stress happens upon a long vowel situated in an open syllable, this long vowel is *overlengthened*.

Moreover, to ensure that the syllabified and stressed results could be shown in the standard Latin and Cyrillic orthographies as well as IPA notation, we adapted a single syllabification function and a single stress function to be applicable across all three systems.

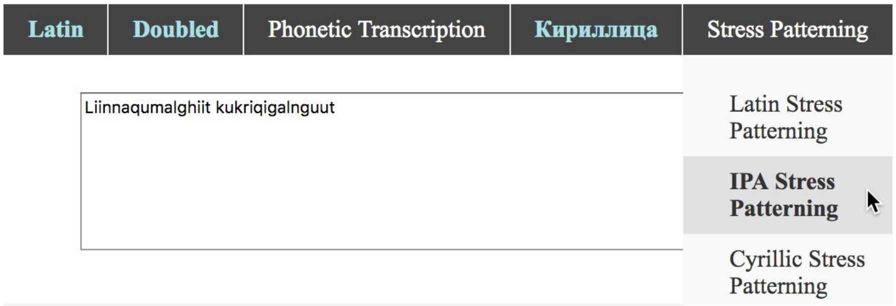
[5] One of these bases is the noun **saaphanghilnguq** “one who can withstand extreme cold”, which we believe is a typo, as both the verb base from which the noun is derived (saapghangite-) and the Cyrillic listing for the noun (**ся̄пҳанӷилңуқ**) use gh /ʁ/ rather than h /h/.

[6] Most of the Yupik bases that include /h/ are particles. For the remaining (non-particle) bases, because our utility does not currently include a morphological analyzer, when transliterating inflected or declined forms of these bases, *г* may be incorrectly transliterated as g. We plan to eventually incorporate a fully automated morphological analyzer as future work, at which point this limitation should be resolved.

4.4 Spell checking Finally, we address the issue of spell checking. Students of Yupik, as well as the fluent adult Yupik population, currently have no tools for automatic spell checking. This lack has also hampered the digitization of existing Yupik corpora, as texts that are scanned and digitized using OCR cannot be automatically checked for errors. While a full spell checker requires an electronic dictionary, we have developed a basic spell checker that relies on Yupik orthographic, phonological, and syllable rules. Specifically, we observe that Yupik restricts syllable structure to CV, CVV, CVC, and CVC (where vowel clusters always consist of a doubled full vowel), and V, VV, VC, and VVC word initially (Krauss 1975, Jacobson 2001). We further rely on the fact that Yupik does not permit diphthongs or geminate consonants (Jacobson 2001).

Our spell checker then simply checks for the presence of illegal sequences:

- CCC clusters
- CC clusters other than those at syllable boundaries
- CC clusters of like consonants
- VV clusters of unlike vowels
- VV clusters where either V is e

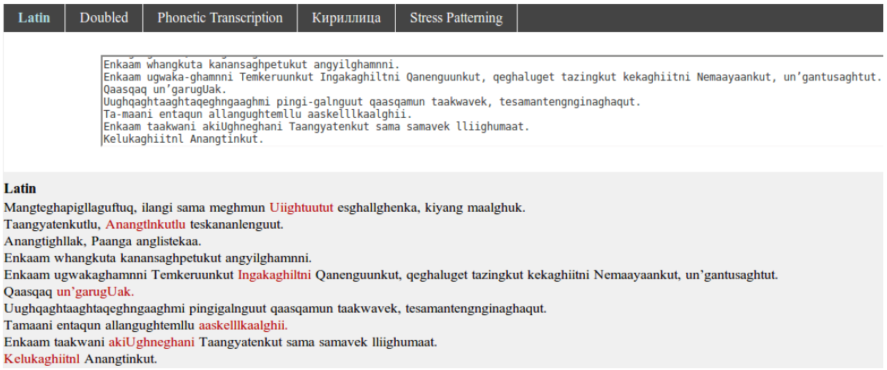


Liinnaqumalghiit kukriqigalnguut

Figure 6. User interface allowing selection from multiple phonetic and stress pattern options.

If a word violates any of these rules, the word is highlighted in red in the output text (see Figures 4 and 7). Despite this spell checker's naivety, it has proven successful in identifying OCR errors in scanned text. Figure 4 shows one such error; the word **tukfiighinaaquteghliaquut** apparently contains the illegal vowel sequence **ia**. A manual examination of the original text, however, reveals that the **i** is in fact an **l** in the original.

In this way, this spell checker has been effective in identifying patterns to OCR errors, in that the letter **l** is often misread as **i** and vice versa, and that unvoiced **ll** is misrecognized as **U** as seen in Figure 7. We expect the inclusion of the spell checker to expedite the process of manually reviewing texts digitized using OCR to produce gold standard documents, and further assist the digitization of Yupik materials.



```
Enkaam whangkuta kanansaghpetukut angyilghamnni.
Enkaam ugwaka-ghamnni Temkeruunkut Ingakaghiltni Qanenguunkut, qeghaluget tazingkut kekaghiitni Nemaayaankut, un’gantusaghtut.
Qaasqaq un’garugUak.
Uughqaghtaaghtaqeghngaaghmi pingi-galnguut qaasqamun taakwavek, tesamantengnginaghaqut.
Ta-maani entaqun allangughtemllu aaskelllkaalghii.
Enkaam taakwani akiUghneghani Taangyatenkut sama samavek lliighumaat.
Kelukaghiitnl Anangtinkut.
```

Latin

Mangteghapigllaguftuq, ilangi sama meghmun Uiightuutut esghallghenka, kiyang maalghuk.
Taangyatenkutlu, Anangtlnkutlu teskananlenguut.
Anangtighllak, Paanga anglistekaa.
Enkaam whangkuta kanansaghpetukut angyilghamnni.
Enkaam ugwakaghamnni Temkeruunkut Ingakaghiltni Qanenguunkut, qeghaluget tazingkut kekaghiitni Nemaayaankut, un’gantusaghtut.
Qaasqaq un’garugUak.
Uughqaghtaaghtaqeghngaaghmi pingigalnguut qaasqamun taakwavek, tesamantengnginaghaqut.
Tamaani entaqun allangughtemllu aaskelllkaalghii.
Enkaam taakwani akiUghneghani Taangyatenkut sama samavek lliighumaat.
Kelukaghiitnl Anangtinkut.

Figure 7. The preliminary spell checker has proven successful in identifying OCR errors in scanned text. The text shown here comes from an OCR-digitized scan of the story *Sivuqam Kiyaghtaallgha Ayumiq* (paragraphs 2–4) by Uvim Ungipaa in Apassingok et al. (1985:10).

5. Conclusion We have developed a set of computational tools which we hope will be of immediate use to members of the Yupik community and to field linguists working with them. Our ultimate goal, in collaboration with the Yupik community, is a full set of tools to enhance the existing bilingual curriculum and provide richer access to existing bilingual materials, adding a digital dimension to language learning that is interactive and dynamic, especially for the younger generations. We also hope that these tools and resources will also enable future documentary linguistic work on this morphologically rich bi-continental language.

References

Apassingok, Anders (Iyaaka) & Edward Tennant (Tengutkalek) (eds.). 1987. *Sulpik*. Unalakleet, Alaska: Bering Strait School District.

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

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

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

Apassingok, Anders (Iyaaka), Willis Walunga (Kepelgu) & Edward Tennant (Tengutkalek), (eds.). 1985. *Sivuqam nangaghnegha – siivanllemta ungipaqellghat / Lore of St. Lawrence Island – Echoes of our Eskimo elders, vol. 1: Gambell*. Unalakleet, Alaska: Bering Strait School District.

Apassingok, Anders (Iyaaka), Willis Walunga (Kepelgu) & Edward Tennant (Tengutkalek) (eds.). 1987. *Sivuqam nangaghnegha – siivanllemta ungipaqellghat*

/ Lore of St. Lawrence Island – Echoes of our Eskimo elders, vol. 1: Savoonga. Unalakleet, Alaska: Bering Strait School District.

Apassingok, Anders (Iyaaka), Willis Walunga (Kepelgu) & Edward Tennant (Tengutkalek), (eds.). 1989. *Sivuqam nangaghnegha – siivanllemta ungipaqellghat / Lore of St. Lawrence Island – Echoes of our Eskimo elders, vol. 1: Southwest Cape*. Unalakleet, Alaska: Bering Strait School District.

Badten, Linda Womkon (Aghnaghaghpik), Vera Oovi Kaneshiro (Uqiıtlek), Marie Oovi (Uvegtu) & Christopher Koonooka (Petuwaq). 2008. *St. Lawrence Island / Siberian Yupik Eskimo dictionary*. Fairbanks, Alaska: Alaska Native Language Center.

Jacobson, Steven A. 2001. *A practical grammar of the St. Lawrence Island/Siberian Yupik Eskimo language*, 2nd ed. Fairbanks, Alaska: Alaska Native Language Center.

Koonooka, Christopher (Petuwaq). 2003. *Ungipaghaghlanga – quutmiit yupigita ungipaghaatangit / Let me tell a story – Legends of the Siberian Eskimos*. Fairbanks, Alaska: Alaska Native Language Center. (Transliterated and translated from the Chukotka collection of G.A. Menovshchikov. Stories told by Ayveghhaq, Tagikaq, Asuya, Alghalek, Nanughhaq, and Wiri.).

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

Krauss, Michael E. 1975. St. Lawrence Island Eskimo phonology and orthography. *Linguistics: An Interdisciplinary Journal of the Language Sciences International Review* 13(152). 39–72.

Krauss, Michael, Gary Holton, Jim Kerr & Colin T. West. 2011. *Indigenous peoples and languages of Alaska*. Fairbanks and Anchorage: Alaska Native Language Center and UAA Institute of Social and Economic Research. ANLC Identifier G961K2010.

Krupnik, Igor. 1983. Early settlements and the demographic history of Asian Eskimos of South Eastern Chukotka (including St. Lawrence Island). In Michael, H. & J. Van Stone (eds.), *Cultures of the Bering Sea region: Papers from international symposium*, 87–111. New York.

Krupnik, Igor. 1994. 'Siberians' in Alaska: The Siberian Eskimo contribution to Alaskan population recoveries, 1880–1940. *Études/Inuit/Studies* 18(1–2). 49–80.

Mailhot, José. 1978. L'étymologie de « Esquimau » revue et corrigée. *Études/Inuit/Studies* 2(2). 59–70.

Nagai, Kayo. 2001. *Mrs. Della Waghiyi's St. Lawrence Island Yupik texts with grammatical analysis*. Endangered Languages of the Pacific Rim A2-006. Kyoto, Japan: Nakanishi Printing.

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

Schwalbe, Daria Morgounova. 2015. Language ideologies at work: Economies of Yupik language maintenance and loss. *Sibirica* 14(3). 1–27.

Shutt, Lauren, Dawn Biddison & Christopher Koonooka. 2014. *Listen & learn: St. Lawrence Island Yupik language and culture video lessons*. Anchorage, Alaska: Arctic Studies Center, Smithsonian Institution.

Tennant, Edward (ed.). 1985. *Yupik formula three reading – Spelling-learning program*. Unalakleet, Alaska: Bering Strait School District.

Vakhtin, Nikolai. 1998. Endangered languages in northeast Siberia: Siberian Yupik and other languages of Chukotka. In Kasten, Erich (ed.), *Bicultural education in the north: Ways of preserving and enhancing indigenous peoples' languages and traditional knowledge*, 159–173. Münster, Germany: Waxmann Verlag.

Vakhtin, Nikolai B. & Evgeniy V. Golovko. 1987. The relations in the Yupik Eskimo sub-group according to lexicostatistics. *Études/Inuit/Studies* 11(1). 3–18.

Lane Schwartz
lanes@illinois.edu
https://orcid.org/0000-0002-3085-2955

Emily Chen
echen41@illinois.edu
https://orcid.org/0000-0003-2609-8133