LANGUAGE DOCUMENTATION & CONSERVATION

Vol. 14 (2020), pp. 69–86
http://nflrc.hawaii.edu/ldc
http://hdl.handle.net/10125/24917
Revised Version Received: 11 Dec 2019

Multidirectional leveraging for computational morphology and language documentation and revitalization

Sylvia L.R. Schreiner
George Mason University

Lane Schwartz
University of Illinois at Urbana-Champaign

Benjamin Hunt
George Mason University

Emily Chen
University of Illinois at Urbana-Champaign

St. Lawrence Island Yupik is an endangered language of the Bering Strait region. In this paper, we describe our work on Yupik jointly leveraging computational morphology and linguistic fieldwork, outlining the multilayer virtuous cycle that we continue to refine in our work to document and build tools for the language. After developing a preliminary morphological analyzer from an existing pedagogical grammar of Yupik, we used it to help analyze new word forms gathered through fieldwork. While in the field, we augmented the analyzer to include insights into the lexicon, phonology, and morphology of the language as they were gained during elicitation sessions and subsequent data analysis. The analyzer and other tools we have developed are improved by a corpus that continues to grow through our digitization and documentation efforts, and the computational tools in turn allow us to improve and speed those same efforts. Through this process, we have successfully identified previously undescribed lexical, morphological, and phonological processes in Yupik while simultaneously increasing the coverage of the morphological analyzer. Given the polysynthetic nature of Yupik, a high-coverage morphological analyzer is a necessary prerequisite for the development of other high-level computational tools that have been requested by the Yupik community.

E-ISSN 1934-5275

1. Introduction

1.1 Overview A number of challenges present themselves in language description and analysis, in the development of computational technologies for under-resourced languages, and in language maintenance and revitalization work.[1] These challenges can be moderated, and community and researcher goals reached more effectively, by leveraging the strengths and tools of each area in support of the others (what we term MULTIDIRECTIONAL LEVERAGING). Resource-building is a goal shared by those working to document languages, build tools for language-learning or language use, increase access to language materials for community use, and improve computational approaches to language analysis. Collective work towards this goal from different angles increases the quality and usefulness of the resources generated and decreases the time needed to produce them.

In this paper we lay out the multilayer virtuous cycle that we have established and continue to refine in our work to document and build tools for St. Lawrence Island Yupik (ISO 639-3 ess, hereafter *Yupik*). After developing a preliminary morphological analyzer from an existing pedagogical grammar of the language, we used it to help analyze new word forms gathered through fieldwork, and improved it through further fieldwork by targeting forms from a corpus of Yupik materials that were not able to be analyzed using the original implementation. The analyzer and other tools we have developed are improved by a corpus that continues to grow through our digitization and documentation efforts, and the computational tools in turn allow us to improve and speed those same efforts.

After a brief description of the language and language situation, we outline the in-tandem operations we are undertaking (§2) and then describe the process and results this approach has yielded in the field (§3). We conclude in §4.

1.2 St. Lawrence Island Yupik Yupik is an endangered language of the Inuit-Yupik-Unangum Tunuu family. It is spoken predominantly in the Bering Strait region, with the majority of speakers residing in communities on St. Lawrence Island and the Chukotka peninsula. Here we briefly discuss a few of the main linguistic features of Yupik, in particular those that present challenges for documentation and computational implementation. Readers with an interest in the grammar of Yupik may wish to consult, for example, Krauss (1975), Jacobson (1977), Jacobson (1985), Krauss et al. (1985), de Reuse (1994), and Jacobson (2001) for more details.

Soviet linguists working in Chukotka in the mid-20th century wrote the first comprehensive descriptions of the Yupik language (Menovshchikov 1960, 1962, 1967; Rubtsova 1971). More recent English-language linguistic work has examined Yupik phonology, prosody, and orthography (Krauss 1975; Jacobson 1985; Krauss et al.

[1] Portions of this work were funded by NSF Documenting Endangered Languages Grants #BCS 1760977 and 1761680, a George Mason University Mathy Junior Faculty Award in the Arts and Humanities, and a George Mason University Presidential Scholarship. Special thanks to the Yupik speakers who have shared their language and culture with us. Thanks also to the audience at ICLDC 2019. We take sole responsibility for any errors.

1985; Jacobson 1990); syntax and language contact (Jacobson 1977, 1994, 2001, 2006; de Reuse 1994); syntax and historical morphology (de Reuse 1992); semantics (de Reuse 2001); morphology and morphophonemics (Vakhtin 2001); polysynthesis (de Reuse 2009); and comparison with Alaskan Yup'ik (Jacobson 2012). The most thorough descriptions are the two-volume Yupik-English dictionary (Badten, et al. 2008) and the pedagogical Yupik grammar (Jacobson 2001).

The documented phonological inventory comprises 31–32[2] consonant and seven vowel phonemes. The consonant inventory includes voiced and voiceless nasals and continuants, and voiceless stops. Vowel phonemes are /i, iː, u, uː, ɑ, ɑː, ə/. There is a relatively simple underlying pattern of alternating stress assignment that is complicated by adjustments that occur when certain vowels are adjacent to each other, and that interacts with vowel length (see especially Krauss 1975). Tautosyllabic consonant clusters are not permitted; syllables are of the form CV(C) or CVV(C) with the first C optional in the first syllable of a word (Jacobson 2005: 9). To our knowledge, allophonic processes have not yet been treated in any published work.

Yupik exhibits ergative-absolutive alignment in its case system and has relatively free word order. The language is pro-drop, and words can generally be categorized by their derivational and inflectional possibilities as nouns, pronouns, or verbs. There is also a three-way system of demonstratives. These vary for case when used pronominally. Descriptors that would be expressed via adjectives in a language like English are generally verbal in Yupik; prepositional relations are expressed via the case system; and adverbial notions are typically conveyed through derivational morphology, particles, and demonstratives. While the derivational morphology is generally agglutinating, the inflectional endings are more fusional. Nominal inflection (on nouns and pronouns) expresses distinctions of case (Jacobson 2001 gives absolutive, "relative" (ergative), ablative-modalis, localis, terminalis, vialis, and equalis); number (singular, dual, and plural); person (first, second, third, and third reflexive, sometimes called "fourth"); nouns but not pronouns also show possession status (unpossessed vs. possessed). Verbal inflection expresses mood (on verbs; Jacobson 2001 lists indicative, participial, interrogative, optative, subordinative, precessive, concessive, consequential, conditional, and contemporative[3]); person (on verbs; first, second, third, and third reflexive, sometimes called "fourth"); transitivity (transitive vs. intransitive); and for transitive verbs, status as subject vs. object. In addition, Yupik is distinguished from its close relatives by its relatively large number of uninflected particles, many of which were borrowed through contact with Chukchi (see de Reuse 1994). In addition to the derivational morphemes with various lexical meanings, Yupik boasts an extensive array of derivational morphemes that signal distinctions that might in other languages be realized as inflections for tense, aspect, voice, and modality.

[2] Jacobson (2001) gives <z> /z/ and <y> /j/ as allophones. However, they are not in perfect complementary distribution. <z> occurs only before <i>, and <y> occurs predominantly but not exclusively before anything other than <i>.

[3] de Reuse (1994: 40–41) does not employ the label "subordinative", and also lists appositional and volitive of fear as moods, splits the contemporative mood into I and II, and splits the participial mood into four: transitive and intransitive participial, and transitive and intransitive participial oblique.

The morphology of Yupik is characterized by extensive agglutination and polysynthesis. Most words in Yupik consist of a nominal or verbal base followed by zero or more derivational suffixes, a single inflectional suffix, and an optional enclitic (of which there are only a handful). There are four types of derivational suffixes ("postbases", in the Yupik literature): noun-elaborating suffixes that affix to nouns and yield nouns, verbalizing suffixes that affix to nouns and yield verbs, verb-elaborating suffixes that affix to verbs and yield verbs, and nominalizing suffixes that affix to verbs and yield nouns. Yupik also exhibits extensive morphophonological changes at morpheme boundaries. Jacobson (2001) documents approximately ten such morphophonological processes that apply when a suffix attaches to a base. Each of these processes is represented explicitly in the grammar and in dictionary entries (Badten et al. 2008) with a unique symbol. For example, the symbol '–' indicates that attaching the suffix to a base will cause the final consonant of that base (if any) to delete. Jacobson (2001) and Badten et al. (2008) treat these processes as lexicalized properties of the suffixes. Under this assumption, the processes that each suffix triggers upon affixation are not environmentally conditioned, and thus cannot be predicted. A small number of non-lexical phonological processes also have been documented, notably vowel dominance, which triggers vowel assimilation to resolve unlike clusters.

Finally, the writing system of the Yupik language reflects its bicontinental history and usage. The Cyrillic orthography was developed several decades earlier than the Latin orthography, and followed a different set of orthographic conventions (Jacobson 1990). The Latin orthography in use on St. Lawrence Island today was developed by linguists working with Yupik speakers in the 1970s. In this orthography, there is a one-to-one correspondence between phonemes and their orthographic representations (graphemes range from one to five characters in length). The orthography does not make use of any diacritics. One particular orthographic pattern is of special interest, as instructors report that it sometimes causes confusion among learners. Of the 31 consonant phonemes, there exist six pairs of continuants and four pairs of nasals that only differ in voicing. In five of the six continuant pairs and in all of the nasal pairs, this difference is marked in the Latin orthography by doubling the grapheme of the voiceless phoneme: for instance, where <m> is the voiced bilabial nasal /m/, <mm> is the voiceless counterpart /m̥/. Orthographic <ll> /ɬ/ and <rr> /ʂ/ also represent the voiceless "counterparts" to <l> /l/ and <r> /ɻ/, respectively, although the segments differ in more than just voicing. The frequency of doubled graphemes naturally lengthens the spelling of words. The Latin orthography exploits the fact that Yupik phonology generally requires adjacent consonants to match in voicing, and allows doubled graphemes to appear as singletons in specific contexts (see Jacobson 2001 and Schwartz & Chen 2017 for details). This orthographic practice is known as UNDOUBLING. We discuss a tool we have developed that addresses this complication in §2.2.

1.3 Speaker population and language ecology Schwartz et al. (2020) estimate that the overall ethnic Yupik population consists of around 2400–2500 native Yupik people (Yupik: *Yupiget*), of whom approximately 800–900 are L1 speakers of Yupik;

this includes approximately 1300 Yupiget on St. Lawrence Island, 800 Yupiget in Chukotka, and 300–400 in mainland Alaska, including Nome and Anchorage. Widespread use of Yupik in Chukotka declined during the mid-20th century (Krupnik & Chlenov 2013; Schwalbe 2017), with estimates of L1 Yupik speakers in Chukotka numbering under 200, all elderly (Vakhtin 2001). Language change is currently underway on St. Lawrence Island; while nearly all St. Lawrence Island Yupiget born prior to 1980 are L1 Yupik speakers (Krauss 1980), Yupik language use among younger generations has dropped precipitously in the succeeding years (Koonooka 2005), with most or all St. Lawrence Island youth today L1 English speakers (Schwartz et al. 2020).

Yupik educational materials were developed by the Bering Strait School District in the 1980s (Tennant 1985; Apassingok et al. 1985, 1987, 1989; Tennant 1989) and 1990s (Apassingok et al. 1993, 1994, 1995). These materials were successfully used in the St. Lawrence Island schools in Gambell and Savoonga during that time period; use of these materials was largely discontinued in the early 2000s in the face of state and federal education mandates coupled with the retirement of experienced Yupik educators (Koonooka 2005). Over the past two years, a community-led language revitalization group has formed in Gambell. During our visits to Gambell during the 2016–2019 timeframe, community leaders in the school, the tribal council, the Native corporation, and the city have expressed to us a strong desire for a robust Yupik language instructional program in the school, with several stating a desire for the eventual introduction of a Yupik immersion program.

2. Building a virtuous cycle Our work with the St. Lawrence Island Yupik community involves three main elements: digitizing materials written in and about Yupik, building tools for language use and education, and further documenting the language. Done in concert, these activities complement each other and increase the benefits obtainable from each, to the advantage of the community, as well as the linguistic and computer sciences. In this section we describe the different levels at which the parts of this process interact, and the benefits gained from these interactions.

2.1 Digitization of legacy materials for corpus building, community access, and tool creation An important first step in supporting the revitalization of Yupik education in St. Lawrence Island schools and in the wider Yupik community is enabling broader access to legacy printed materials in and about Yupik. This is accomplished by digitizing these documents and making the resulting electronic documents available to the St. Lawrence Island schools and community members.

Our first priority was to start digitizing Yupik-language texts that would be of use to the community and would also help build a Yupik-language corpus for use in education, language revitalization, and research. Substantial amounts of monolingual Yupik and bilingual Yupik-Russian written materials were developed in Chukotka in the 1930s–1950s, primarily for pedagogical use (Krauss 1973). During the 1970s–1990s, monolingual Yupik and bilingual Yupik-English written materials were developed in Alaska by the Bureau of Indian Affairs, the University of Alaska, and the Bering Strait School District, also primarily for pedagogical use (see De Reuse 1994 and Schwartz et al. 2020 for more details). We began with the three-volume *Lore of St. Lawrence Island* (Apassingok et al. 1985; 1987; 1989), a set of bilingual Yupik-English elementary readers (Apassingok et al. 1993; 1994; 1995), and a collection of stories from Chukotka (Koonooka 2003). Prior to our digitization work, these important Yupik corpora existed in paper format only.

Our second priority has been the pedagogical and reading materials located in the Gambell School library and Materials Development Center. This includes approximately 90 elementary-level primers (e.g., Apassingok & Waghiyi 1985a; 1985b) and the Yupik bilingual-bicultural curricula (Tennant 1989). All of these materials have been scanned; we are in the process of processing the raw scanned images into accessible PDFs with embedded searchable text. Several copies of some of the primers exist, but to our knowledge there is only one copy of almost all of the curricular materials, in paper form in three-ring binders filling several shelves. These materials are not currently in use at the school. These digital materials will be accessible by educators and, when appropriate, students and the greater Yupik community, to support education and revitalization efforts. We are also working closely with archivists at the Alaska Native Language Archive (ANLA) in Fairbanks, Alaska to catalog all of the elementary-level primers that have not yet been assigned a permanent identifier through the archive. In this way, the ANLA is one means for long-term preservation of existing materials.

There is a substantial collection of Yupik language materials and academic work on Yupik at the ANLA. Much of this material exists only in paper form, at the Archive. We are working to digitize all of these items, and in doing so, assist the Archive in indexing several crates and boxes of uncatalogued Yupik materials as well. As of March 2019, approximately 60 percent of the Archive's indexed Yupik materials have been scanned and assembled for additional cleanup and processing. These include a considerable number of folders of loose-leaf field notes, Kayo Nagai's (2004) dissertation; elementary readers; two volumes of St. Lawrence Island folk tales and legends; six books of the New Testament in both the Latin and Cyrillic orthographies; and a book of hymns, again in both orthographies.

In addition to the materials that have a more direct use for the community, such as the readers and religious texts, we are also digitizing the interlinearly glossed examples from existing scholarly work. These examples increase the size of the unglossed Yupik corpus and form the beginnings of a glossed corpus. These include the entirety of Nagai's (2001) collection of texts with analysis, as well as the examples found in de Reuse's (1994) work and the other academic works on Yupik that include interlinear glossing (e.g. Vakhtin 1989; de Reuse 1992; Vakhtin 2000; Nagai 2004; de Reuse 2009).

Digitizing the materials involves scanning each page of the resource on a flatbed scanner with a sloped edge as a TIFF image at 600 dpi. Subsequent cleanup utilizes the open source ScanTailor software to correctly orient pages, deskew, despeckle, dewarp, and center the selected content of each image. The final TIFF images are converted to PDF format and merged into a single file in Adobe Acrobat Pro XI.

Several factors drove our decision to make digitization an early and significant part of the effort to support the community's language maintenance and revitalization goals. One major driving force is accessibility. Most materials were only accessible on-site at the Gambell School or the Alaska Native Language Archive. As we digitize materials that are in Yupik, they add to a growing searchable corpus of the language. Throughout this process, we have consulted with the respective copyright holders and with representatives of the St. Lawrence Island Yupik community, and have received permission to distribute digitized corpora under a Creative Commons Attribution Non-Commercial 4.0 International License. In collaboration with the Alaska Native Language Archive, we will index and archive digitized materials and ensure that each item is assigned a permanent archival identifier. We will make these materials accessible digitally to Yupik community members who want to read the stories that have been collected or access other materials. This corpus also has potential benefits for researchers, including linguists studying the language.

In addition to direct benefits to the community, digitization fits into the virtuous cycle being described here in establishing an electronic corpus of Yupik. Computational linguistic research on the Yupik language requires Yupik text in electronic form. A digitized corpus of Yupik texts directly enables computational linguists to test models of the Yupik language. Our finite-state morphological analyzer (described in §2.2) models the lexicon and morphophonology of Yupik, following Jacobson's (2001) description of the language. As we digitize Yupik texts into accessible electronic form, we have been able to test the morphological analyzer against the digitized texts. This process exposes lexical entries not in the Yupik dictionary as well as instances in the corpus where Jacobson's (2001) grammar fails to account for linguistic phenomena extant in the data. The addition of the interlinear glossed corpus will help us further improve the finite-state morphological analyzer and will also benefit linguists interested in Yupik or related languages, or specific linguistic phenomena. Beyond the benefits provided by the existence of a more extensive corpus, some of our digitization efforts feed directly into tools that community members can use for language use or teaching. Digitization of pedagogical materials provides an immediate benefit for teachers working to add to or improve existing curricula, allowing them to print premade worksheets and other materials, or modify them for new activities. These will also be of use in preparing an eventual immersion curriculum. With the picture-filled elementary primers, we are creating a collection of e-books with audio narration in Yupik, which can be used in the school or in the homes of community members. We further discuss these tools in the following subsection.

2.2 Building tools for language use and analysis The second element of our approach is the development of computational tools for language analysis, learning, and use. We view computational tool development as integral to language documentation and revitalization. With respect to language documentation, the development and availability of tools such as a morphological analyzer and resources such as digitized texts serve a joint purpose, increasing a field linguist's research productivity while simultaneously enabling and encouraging the linguist to empirically evaluate hypotheses against a much larger dataset than could be practically consulted as part of manual analysis. With respect to language revitalization, the availability of computational language technologies can provide direct benefits, for example in educational settings, as well as indirect benefits, such as serving to increase a language's perceived status in the eyes of community youth.

Some of the tools we are developing are directly usable by the community, while others form part of a larger technological infrastructure that will support research towards the future creation of more technically sophisticated tools. To date, we have developed a suite of web-based orthographic utilities (Schwartz & Chen 2017), a finite-state morphological analyzer (Chen & Schwartz 2018), a preliminary neural morphological analyzer (Schwartz et al. 2019a), and an electronic dictionary (Hunt et al. 2019).[4] We are also creating digital language utilities such as interactive, narrated e-books (from the printed elementary primers, Schwartz et al. 2019b) and a "read-along" video of the Pledge of Allegiance in Yupik (at the request of teachers at the school, who reported that younger students struggle with the Yupik words).

Of the orthographic utilities we have developed, some are intended for use by the community, while others are meant primarily for researchers. The most basic utility performs as a non-lexical, orthotactic spellchecker.[5] This allows us to catch most common OCR errors during the digitization process prior to human intervention. One of the utilities transliterates from the Yupik Latin orthography to any of several phonetic orthographies, with an option to include stress marking. This is of more use to researchers beginning to work with the language, but might eventually also be co-opted for use in language learning environments. Another utility transliterates between the Latin and Cyrillic orthographies for the language. This utility will be integrated into the corpus of Yupik-language materials, so that speakers from Chukotka can easily access the corpus as well. We also anticipate this being useful for possible sharing of pedagogical materials that are developed on St. Lawrence Island with educators in Russia. We developed another utility at the request of several teachers at the Gambell school, who related to us that students coming in with limited Yupik speaking ability were having trouble getting used to the "undoubling" convention in the standard orthography (as mentioned in §1.2). In Yupik, if a voiced and an unvoiced consonant appear next to each other in a word, the voiced phoneme will almost always devoice to assimilate.[6] Typically, the voiceless counterpart of a voiced sound is represented orthographically with a doubled letter (e.g., <g> represents /ɣ/ and <gg> represents /x/). The orthography takes advantage of the fact that speakers of Yupik will know the devoicing rule subconsciously and "undoubles" these two-letter sequences when they are predictable due to the phonological rule, resulting in a shorter overall word-

[4] The orthographic tools can be found and utilized here: https://saintlawrenceislandyupik.github.io/web_tools/index_utilities.html, and the dictionary here: https://saintlawrenceislandyupik.github.io/web_tools/index_dictionary_transducer.html, The source code for the tools and dictionary can be found at https://github.com/SaintLawrenceIslandYupik/web_tools, and the source code for the morphological analyzer can be found at https://github.com/SaintLawrenceIslandYupik/finite_state_morphology.

[5] That is, a spellchecker that checks for violations of the patterns of and restrictions on graphemes in orthographic representations of words of the language. This tool and the transliteration utility are each implemented in Python for offline use and in Javascript for use in web-based utilities.

[6] An exception is that voiced nasals may be followed by unvoiced consonants (Jacobson 2001:5).

form. For example, a word including the phonemic sequence /g/+/m/ will be written <gm>, while one with the sounds /k/+/m/ will be pronounced [km̥] but written <km>, rather than the otherwise expected <kmm>. When students encounter <km> in a new word, they reportedly will often pronounce it [km] rather than the correct [km̥]. Our "doubling" utility takes standard Yupik orthography as input and returns a version with undoubled letters restored. The hope is that students may find it useful to toggle between outputs to help learn the patterns.

A morphological analyzer is a good example of a tool whose usefulness may not be immediately apparent to community members, but which forms the basis of a number of other utilities benefitting speakers as well as researchers. Our existing Yupik morphological analyzer is implemented as a finite-state transducer and was developed in the *foma* finite-state toolkit (Hulden 2009). It is a faithful adaptation of the grammatical and morphophonological rules described in Jacobson (2001), and encompasses all of the lexical items listed in the Badten (2008) dictionary, including approximately 500 particles, 4000 noun roots, 4000 verb roots, and 600 derivational suffixes. While implementation of the finite-state analyzer demanded several months of dedicated effort, its central design can be readily adapted for Yupik's sister language, Alaskan Yup'ik (ISO 639-3 esu). Moreover, since the *foma* toolkit supports porting of finite-state transducers to other formats, specifically a Javascript object, we were able to swiftly integrate our finite-state analyzer into our preliminary electronic dictionary so that users of the dictionary do not need to know the underlying or base form of a word in order to look it up. An accurate morphological analyzer is also needed for certain resources requested by the community, such as a spellchecker and text completion. For the researcher, the analyzer also aids in corpus searches, yielding faster and more accurate results when searching for a particular morpheme or base word that may not ever match an expected citation or base form within a corpus. We further discuss the cooperative field use of the analyzers to improve (and be improved by) fieldwork in §3.

2.3 Data collection for language analysis and resource improvement Along with digitization and tool-building, a major goal of our fieldwork is the documentation of Yupik phonology, morphology, and syntax beyond that described by Krauss (1975) and Jacobson (2001). In addition to the clear benefits that this documentation can have for linguistic science, it also adds data to the growing corpus. We hope that our continued documentation will also be of use for developing more modern pedagogical materials for Yupik language instruction and immersion programs. Significantly, this documentation also helps us improve the tools we are building: filling in gaps in what we (as non-speakers) know of the language in order to improve the finite state morphological analyzer, add words and phrases to future initial versions of language learning applications, etc. We address this part of the virtuous cycle further in §3.

We work with speakers individually or in pairs or groups, according to their preference, and combine naturalistic and semi-naturalistic production with targeted elicitation (as much as possible using Yupik forms that we have pulled from existing materials or generated, rather than translating from English to Yupik). Targeted elicitation, along with detailed positional and semantic work with derivational morphology, is used to address particular underdocumented morphosyntactic or semantic phenomena, as well as insufficiencies in the finite-state analyzer (as discussed in the following section). We have also worked with several speakers on word-by-word (or phrase-by-phrase) translations of Yupik texts that otherwise only have more free translations. These will form the basis of interlinear glosses for the corpus, which we will undertake with the help of the morphological analyzers.

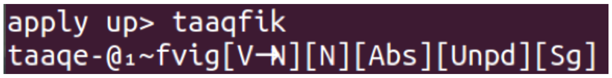
Our current priorities for documentation reflect the collaborative nature of our work. For the sake of improving the documentation of Yupik, we are prioritizing un(der)documented syntactic and morphological phenomena, including working to establish a better understanding of the (morpho)syntactic functional hierarchy; clarifying the workings of phenomena for which there is conflicting information in the existing literature; and documenting allophonic variation in vowels and consonants. Then, with respect to our computational goals, our priorities are in clarifying und(der)documented morpheme attachment rules, lexical items, orthographic variants, etc. that we hypothesize to underlie many if not all of the errors returned by our finite-state morphological analyzer.

3. Multidirectional leveraging in the field In this section we describe in detail a significant portion of the virtuous cycle, that between the finite-state morphological analyzer and the fieldworker. While the orthography is at least phonemically transparent, Yupik's large number of derivational and inflectional suffixes coupled with varying rule-based but complex processes for morphophonological attachment make on-the-spot analysis of new word forms difficult for non-speakers or learners. This combination of linguistic features also makes corpus searching less than straightforward. Many derivational suffixes have a number of instantiations, depending on the phonology of the base they attach to in a given word form. A working morphological parser helps on both fronts. Elicitors are aided during the data collection process by using the parser to analyze new data. This new data in turn helps to improve the parser. In the other direction, word forms that cause the parser to return an error can be used to direct fieldwork towards holes in existing documentation. Answering the questions posed by parser errors then further improves the finite-state analyzer.

3.1 Enhancing elicitation and analysis Sometimes, a team member working with a speaker or speakers encounters a word form and finds it difficult to parse manually (or wishes to parse it more quickly than they would be able to by hand). In this case, the user inputs the form and one of three things happens.

First, the morphological analyzer may return a single analysis. For example:

Figure 1. The analyzer returns a single analysis

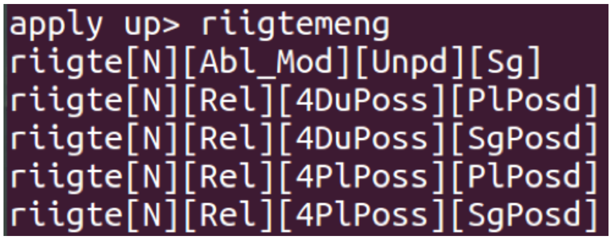


```
apply up> taaqfik
taaqe-@₁~fvig[V→N][N][Abs][Unpd][Sg]
```

In this instance, the user is presented with a single option for analysis of the word form. That is, they are presented with a single base word and one combination of derivational and inflectional endings. There are several possible situations that can lead to this outcome. First (and hopefully the case), the analysis may be correct, and the only possible correct analysis. Second, the given analysis may be correct, but there may be other possible analyses that the parser has missed. Third, the given analysis may be incorrect (that is, it is not actually a possible analysis), and the parser has missed the correct possible analysis or analyses. In each of these cases, the team member can follow up with the speaker (and/or another speaker) for help in confirming the analysis. It is more difficult to confirm that there are no other possible correct analyses, but some speaker responses can help on this point. For instance, if the speaker identifies another meaning for the word form in question that cannot be understood to be yielded from the analysis under scrutiny, there may be an analyzer error. To give an overly simple example: if the analyzer produces a single analysis with third person singular inflection, and the user or a speaker knows that the word form can also refer to third person plural, then the analyzer has failed to produce all possible analyses.

Second, the analyzer may return multiple analyses.

Figure 2. The analyzer returns multiple analyses

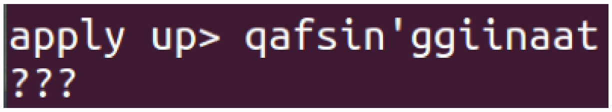


```
apply up> riigtemeng
riigte[N][Abl_Mod][Unpd][Sg]
riigte[N][Rel][4DuPoss][PlPosd]
riigte[N][Rel][4DuPoss][SgPosd]
riigte[N][Rel][4PlPoss][PlPosd]
riigte[N][Rel][4PlPoss][SgPosd]
```

Here, the user is presented with a number of different possible breakdowns of the word form. Again, more than one path can lead to this outcome. The analyzer may have yielded all the correct analyses (the aim), or it may have missed some correct analyses. (Of course, it may have done either of these and also yielded some incorrect analyses.) In any of these cases, the user can narrow possibilities based on meaning and context as far as possible, and they can return to the speaker with further questions to help determine the correct analysis for the context. Sometimes, a user may want to know all possible decompositions of a given word form. However, depending on the circumstances of use, a user may prefer to have the analyzer return a single, "most likely" analysis. Because the current analyzer is built on a non-probabilistic finite-state transducer, it is unfortunately unable to provide a probabilistic ranking of returned analyses. Developing a reliable probabilistic morphological analyzer capable of returning ranked analyses is a matter of ongoing research.

Third, the analyzer may return no analysis.

Figure 3. The analyzer returns no analysis



```
apply up> qafsin'ggiinaat
???
```

In this case, the analyzer has been unable to determine an analysis based on the input. The possible reasons for this are discussed in the next section.

In all cases, the analyzer helps the user more quickly narrow down the possibilities for on-the-spot analysis and better understanding in the field. The analyzer can be consulted ("up") during an elicitation session to help the user more quickly analyze word forms about which they are unsure of the morphological decomposition, and can also be applied "down" to yield a Yupik word form from the desired component parts, more quickly and accurately. The user can also consult the analyzer between sessions, to help guide the next one.

Interlinear glossing is important in corpus and published academic work; particularly for scholars but also for learners and speakers interested in approaching their language in a different manner. Regardless of whether a form is analyzed entirely by a researcher or whether the analyzer is used, interlinear glossing will continue to be important so long as it is the standard for field notes, corpora, and academic publications. In addition, already-analyzed and glossed materials that form part of the growing corpus of Yupik will help train the neural analyzer that is currently in progress.

A perfectly accurate analyzer would make a good deal of by-hand analysis unnecessary. However, even an analyzer that gave error-free analyses would still not be able to choose between competing accurate analyses–that is, in cases for which there is more than one possible analysis of the form (reflecting syncretism(s) in a paradigm). Those cases will still require decisions by the researcher. Of course, any morphological analyzer will require extensive testing and proving before it can be trusted to accomplish more than a first pass at analysis.

In the current version of the analyzer, input in Yupik is equivalent to the top line of a three-line gloss, and the output is essentially the morpheme-by-morpheme gloss line (its formatting does not match what you would typically see in an academic paper, as can be seen in figures 1–3, but it would be relatively trivial to adjust this). What is missing from the analyzer currently is the free translation into English. Machine translation of this type is one of our long-term goals (particularly English-to-Yupik translation, given the goals of the community to have more reading and educational materials accessible in Yupik) but will require extensive additional work, especially given the polysynthetic nature of the language and the relatively sparse existing documentation of the meanings (English translations, usage notes, fine-grain distinctions between meanings, etc.) of the many derivational morphemes, particularly when occurring in combination with other morphemes. As the Yupik corpus grows, we will be able to work towards reliable machine translation.

3.2 Improving the analyzer The analyzer may also fail to process a form drawn from our existing corpus or elicited data, or fail to produce what we know to be the correct analysis. By hypothesis, a form that triggers such an output probably involves one or more orthographic, phonological, morphological, or lexical phenomena (or variations) that are currently undocumented or not well-documented. (Less interesting reasons include errors in input and errors in the analyzer code, which must be ruled out first.)

In our work so far, we have found errors due to issues in each of these areas. Orthographic variation causes some errors. There are several causes of such variation. There are some speaker-to-speaker individual differences in spelling, as found in any speech community. This is noticeably aggravated in an otherwise fairly transparent orthography by the orthographic undoubling described in §1. For instance, some speakers do not consistently adhere to this orthographic convention, instead writing words in their doubled form, such as *naa**llk**enaaghaat* rather than *naa**lk**enaaghaat*. There are also spelling variations depending on the family and/or clan of the speaker. Thus, the dictionary-standard orthographic representation of a given word may not be the accepted or usual spelling for all individuals.

We have also found previously undescribed phonological processes that operate across word boundaries, rendering the surface form opaque. One such phonological process can be represented by the rule /t/ → [s] // _ # t, where '#' designates a word boundary. For example, the Yupik word *aatqus* in the excerpted phrase *aatqus tamaakut* (Apassingok et al. 1985) is, in fact, underlyingly *aatqut*. In this context, it has been phonologically (and orthographically) altered to prevent consonant gemination across word boundaries. Other phonological rules of this type include <k> /k/ → <q> /q/ and <q> /q/ → <gh> /ʁ/, although the conditions under which they occur are not well-studied.

Phonological variation also plays a role in orthographic variation. For instance, vowel length is phonemic and represented orthographically, but there is variation among speakers as to vowel length in certain words. This then can lead to orthographic variation if the speaker chooses not to use, or is unaware of, the dictionary-standard spelling. The Yupik word for 'flower', for instance, is spelled *piitesighaq* per the Badten (2008) dictionary, but an alternative spelling, *piitesiighaq* with a doubled second <i>, can be found in Nagai (2001) (and both pronunciations are heard).

Yupik is polysynthetic and boasts over 600 derivational morphemes, each of which attaches to the preceding morpheme or base according to one of a number of semi-regular morphophonological processes. Other parser errors have been caused both by previously undescribed morphemes, and by previously undescribed, or incompletely described, affixation processes or environments. For instance, in conducting fieldwork we found that the Jacobson (2001) reference grammar does not account for a form of the third person singular intransitive contemporative mood that appears frequently in the *Lore of St. Lawrence Island* trilogy (Apassingok et al. 1985; 1987; 1989). Whereas Jacobson (2001) gives this inflectional ending as *-neghani*, it often appears as *-n'ghani* or alternatively, *-n'ghaani* in the trilogy instead, as in *aaskestaaghhaan'ghani* and *aliin'ghun'ghaani* (Apassingok et al. 1985), respectively.

We have also come across several previously unrecorded or undescribed lexical items. For instance, the verbal suffix *-kestaamaan* 'in the time of' does not appear in either the Jacobson (2001) reference grammar or the Badten (2008) dictionary. It was identified by two of the speakers we worked with in several words that the analyzer could not parse. In addition to improving the analyzer, these items can also be easily added to the digitized version of the dictionary we have created if the community wishes.

Once each of these issues was identified, the analyzer and/or lexicon were adjusted to reflect the new knowledge. Previously undocumented morphemes and lexical items were simply added to the lexicon of the analyzer, while pre-processing steps to convert words to their underlying forms were implemented to account for the cross-word boundary phonological processes. Spelling alternations have not yet been explicitly handled, however, given the many variations in spelling that may occur. We intend to eventually introduce a feature that attempts to analyze known spelling variations of a word if the original input word cannot be analyzed.

Fieldwork helps improve the accuracy of the analyzer in two ways. One, unanalyzable forms from the existing corpus are made analyzable by amending the source code to reflect new knowledge about the orthography, phonology, morphology, or lexicon garnered from elicitations. Two, texts recorded via fieldwork (stories, sentences, etc.) can be added to the corpus and provide further testing of the analyzer, and errors can be investigated later through elicitation. The analyzer, in turn, helps to improve our record of the language by producing errors in response to unanalyzable forms which serve to identify areas in which the record is under- or misinformed.

Augmentations to the finite-state analyzer during piloting of this process in Summer 2018 led to a reduction in unanalyzed word types from 25 percent to 22 percent of word types. This corresponds to approximately 2000 previously unanalyzed Yupik tokens that received at least one analysis from the analyzer where they had previously received none. So far, the errors we have investigated through elicitation have come primarily from running the analyzer over the existing corpus. However, the corpus currently only consists of approximately 30,000 unique words or types. This comprises only a fraction of all Yupik word forms, given the polysynthetic nature of its word-building strategies. As the corpus grows, we expect to continue to find more parser errors, which will in turn help us identify more gaps in the documentation. As we address those errors, the parser's coverage will improve, and the errors will decrease accordingly.

4. Conclusion

Establishing a virtuous cycle among the processes of digitization, tool-building, and documentation benefits all sides of the research process. Digitization of existing materials builds a corpus that can be of direct use to the community for language maintenance, learning, and revitalization, and is necessary for the improvement of morphological analyzer(s). A reliable, automated morphological analyzer of Yupik presents a number of benefits. In addition to helping speed the documentation process by aiding analysis and pointing out holes in existing documentation, it facilitates development of computational resources for the community, including projects further downstream such as speech recognition and machine translation. In addition to improving the analyzer, further documentation yields a better record of the language in the form of a larger corpus and an improved picture of the workings of the phonology, morphology, and syntax of Yupik.

The Yupik community's goals of establishing an immersion curriculum and, even before that, using more Yupik-language materials in the classroom, are supported by efforts to digitize existing pedagogical materials, and by the establishment of an easily accessible and searchable corpus. We hope that the improvement of existing resources and the increased availability thereof will significantly benefit the revitalization goals and efforts of the speaker community.

Those speaking, learning, or documenting other languages in the family may benefit from the improved documentation of Yupik and subsequent computational tools that come with this process. This process may also be of use to others working on underdocumented languages, particularly those that make use of complex wordforms.

References

Apassingok, Anders (Iyaaka), Willis (Kepelgu) Walunga, & Edward (Tengutkalek) Tennant (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. http://www.uaf.edu/anla/collections/search/resultDetail.xml?id=SY980AWT1985.

Apassingok, Anders (Iyaaka), Willis (Kepelgu) Walunga, & Edward (Tengutkalek) Tennant (eds.). 1987. *Sivuqam nangaghnegha – Siivanllemta ungipaqellghat / Lore of St. Lawrence Island – Echoes of our Eskimo elders, vol. 2: Savoonga*. Unalakleet, Alaska: Bering Strait School District. http://www.uaf.edu/anla/collections/search/resultDetail.xml?id=SY980AWT1985.

Apassingok, Anders (Iyaaka), Willis (Kepelgu) Walunga, & Edward (Tengutkalek) Tennant (eds.). 1989. *Sivuqam nangaghnegha – Siivanllemta ungipaqellghat / Lore of St. Lawrence Island – Echoes of our Eskimo elders, vol. 3: Southwest Cape*. Unalakleet, Alaska: Bering Strait School District. http://www.uaf.edu/anla/collections/search/resultDetail.xml?id=SY980AWT1985.

Apassingok, Anders (Iyaaka), Jessie (Ayuqliq) Uglowook, Lorena (Inyiyngaawen) Koonooka, & Edward (Tengutkalek) Tennant (eds.). 1993. *Kallagneghet / Drumbeats. Unalakleet, Alaska: Bering Strait School District*. http://www.uaf.edu/anla/item.xml?id=SY990AUKT1993.

Apassingok, Anders (Iyaaka), Jessie (Ayuqliq) Uglowook, Lorena (Inyiyngaawen) Koonooka, & Edward (Tengutkalek) Tennant (eds.). 1994. *Akiingqwaghneghet / Echoes*. Unalakleet, Alaska: Bering Strait School District. http://www.uaf.edu/anla/item.xml?id=SY990AUKT1994.

Apassingok, Anders (Iyaaka), Jessie (Ayuqliq) Uglowook, Lorena (Inyiyngaawen) Koonooka, & Edward (Tengutkalek) Tennant (eds.). 1995. *Suluwet / Whisper-*

ings. Unalakleet, Alaska: Bering Strait School District. http://www.uaf.edu/anla/item.xml?id=SY900AUKT1995.

Apassingok, Anders (Iyaaka) & Dorothy Waghiyi. 1985a. *Talli*. Gambell, Alaska: St. Lawrence Island Bilingual Education Center.

Apassingok, Anders (Iyaaka) & Dorothy Waghiyi. 1985b. *Qula*. Gambell, Alaska: St. Lawrence Island Bilingual Education Center.

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

Chen, Emily & Lane Schwartz. 2018. A morphological analyzer for St. Lawrence Island / Central Siberian Yupik. In *Proceedings of the 11th Language Resources and Evaluation Conference (LREC-2018)*, 2623–2630. https://www.aclweb.org/anthology/L18-1416.

de Reuse, Willem J. 1992. The role of internal syntax in the historical morphology of Eskimo. In Aronoff, Mark (ed.), *Morphology now*, 163–178. Albany, New York: SUNY Press.

de Reuse, Willem J. 1994. *Siberian Yupik Eskimo: The language and its contacts with Chukchi. Studies in indigenous languages of the Americas*. Salt Lake City, Utah: University of Utah Press.

de Reuse, Willem J. 2001. The great Yupik mood swing, and its implications for the directionality of semantic change. In Adronis, Mary (ed.), *CLS 37: The panels*, 239–247. Chicago: Chicago Linguistic Society.

de Reuse, Willem J. 2009. Polysynthesis as a typological feature: An attempt at a characterization from Eskimo and Athabascan perspectives. In Mahieu, Marc-Antoine & Nicole Tersis (eds.), *Variations on polysynthesis: The Eskaleut languages*, 19–34. Amsterdam: John Benjamins.

Hulden, Mans. 2009. Foma: A finite-state compiler and library. In *Proceedings of the 12th conference of the European chapter of the Association for Computational Linguistics demonstrations session*, 29–32. Stroudsburg, Pennsylvania: Association for Computational Linguistics.

Hunt, Benjamin, Emily Chen, Sylvia L.R. Schreiner, & Lane Schwartz. 2019. Community lexical access for an endangered polysynthetic language: An electronic dictionary for St. Lawrence Island Yupik. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, 122–126. https://www.aclweb.org/anthology/papers/N/N19/N19-4021/.

Jacobson, Steven A. 1977. *A grammatical sketch of Siberian Yupik Eskimo as spoken on St. Lawrence Island, Alaska*. Fairbanks, Alaska: Alaska Native Language Center.

Jacobson, Steven A. 1985. Siberian and Central Yupik prosody. In Krauss, Michael E. (ed.), *Yupik Eskimo prosodic systems: Descriptive and comparative studies*, 25–46. Fairbanks, Alaska: Alaska Native Language Center.

Jacobson, Steven A. 1990. Reading and writing the Cyrillic System for Siberian Yupik. Fairbanks: Alaska Native Language Center. http://www.uaf.edu/anla/item.xml?id=SY975J1990a.

Jacobson, Steven A. 1994. The 'observational construction' in Central (Alaskan) Yup'ik Eskimo and (Central) Siberian Yupik Eskimo. *Acta Linguistica Hafniensia* 27. 261–274.

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

Jacobson, Steven A. 2005. History of the Naukan Yupik Eskimo dictionary with implications for a future Siberian Yupik dictionary. *Études Inuit Studies* 29(1–2). 149–161.

Koonooka, Christopher (Petuwaq). 2003. *Ungipaghaghlanga – Quutmiit Yupigita ungipaghaatangit / Let Me Tell a Story – Legends of the Siberian Eskimos*. Fairbanks, Alaska: Alaska Native Language Center.

Krauss, Michael E. 1973. *St. Lawrence Island and Siberian Eskimo Literature*. http://www.uaf.edu/anla/item.xml?id=SY970K1973a.

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

Krauss, Michael E., Jeff Leer, Steven A. Jacobson, & Lawrence Kaplan (eds.). 1985. *Yupik Eskimo prosodic systems: Descriptive and comparative studies*. Fairbanks, Alaska: University of Alaska Press.

Krupnik, Igor & Michael Chlenov. 2013. *Yupik transitions: Change and survival at Bering Strait, 1900–1960*. Fairbanks, Alaska: University of Alaska Press.

Menovshchikov, G. A. 1960. *Eskimosskii iazyk*. Leningrad: Gosudarstvennoe uchebno-pedagogicheskoe izdatel'stvo.

Menovshchikov, G. A. 1962. *Grammatika iazyka aziatskikh eskimosov, vol. 1*. Moscow and Leningrad: Izdatel'stvo akademii Nauk.

Menovshchikov, G. A. 1967. *Grammatika iazyka aziatskikh eskimosov, vol. 2*. Moscow and Leningrad: Izdatel'stvo akademii Nauk.

Menovshchikov, G. A. & Nicolai B. Vakhtin. 1990. *Eskimosskii iazyk*. Leningrad: Prosveshchenie. Preliminary edition 1983.

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*. Kyoto, Japan: Nakanishi Printing.

Nagai, Kayo. 2004. *A morphological study of St. Lawrence Island Yupik: Three topics on referentiality*. Kyoto, Japan: Kyoto University. (Doctoral dissertation.)

Rubtsova, E. C. 1971. *Eskimossko-russkii slovar'*. Moscos: Izdatel'stvo "Sovetskaia entsiklopediia".

Schwalbe, Daria Morgounova. 2017. Sustaining linguistic continuity in the Beringia: Examining language shift and comparing ideas of sustainability in two Arctic communities. *Anthropologica* 59(1). 28–43. https://muse.jhu.edu/ article/658679.

Schwartz, Lane & Emily Chen. 2017. Liinnaqumalghiit: A web-based tool for addressing orthographic transparency in St. Lawrence Island/Central Siberian Yupik. *Language Documentation & Conservation* 11. 275–288. http://hdl.handle.net/10125/24736.

Schwartz, Lane, Emily Chen, Benjamin Hunt, & Sylvia L.R. Schreiner. 2019a. Bootstrapping a neural morphological analyzer for St. Lawrence Island Yupik from a finite-state transducer. In *Proceedings of the 3rd Workshop on Computational Methods for Endangered Languages: Vol. 1, Article 12*, 85–96. https://scholar.colorado.edu/scil-cmel/vol1/iss1/12.

Schwartz, Lane, Sylvia L.R. Schreiner, Peter Zukerman, Giulia Masella Soldati, Emily Chen, & Benjamin Hunt. 2019b. Initiating a tool-building infrastructure for the use of the St. Lawrence Island Yupik language community. Talk, International Year of Indigenous Languages 2019: Perspectives Conference, October 2019.

Schwartz, Lane, Sylvia L.R. Schreiner, & Emily Chen. 2020 (In press). Community-focused language documentation in support of language education and revitalization for St. Lawrence Island Yupik. *Études Inuit Studies* (Forthcoming).

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

Tennant, Edward (ed.). 1989. *Yupik language and culture curriculum*. Unalakleet, Alaska: Bering Strait School District.

Vakhtin, Nikolai B. 1989. Towards order analysis of Yupik Eskimo verbal inflection. *Études/Inuit/Studies* 13(1). 115–130.

Vakhtin, Nikolai B. 2001. *Yazyki narodov Severa v XX veke: Ocherki yazykovogo sdviga*. St. Petersburg, Russia: European University at St. Petersburg.

Sylvia L.R. Schreiner
sschrei2@gmu.edu
orcid.org/0000-0003-2394-3477

Lane Schwartz
lanes@illinois.edu
orcid.org/0000-0003-2609-8133

Benjamin Hunt
bhunt6@gmu.edu
orcid.org/0000-0001-5536-3026

Emily Chen
echen41@illinois.edu
orcid.org/0000-0002-3085-2955