NRC·CNRC

NRC Publications Archive
Archives des publications du CNRC

Kawenn´on:nis: the Wordmaker for Kanyen'k´eha

Kazantseva, Anna; Maracle, Owennatekha Brian; Maracle, Ronkwe'tiyóhstha Josiah; Pine, Aidan

This publication could be one of several versions: author's original, accepted manuscript or the publisher's version. / La version de cette publication peut être l'une des suivantes : la version prépublication de l'auteur, la version acceptée du manuscrit ou la version de l'éditeur.

Publisher's version / Version de l'éditeur:

Proceedings of the Workshop on Computational Modeling of Polysynthetic Languages, pp. 53-64, 2018-08-26

NRC Publications Record / Notice d'Archives des publications de CNRC:
https://nrc-publications.canada.ca/eng/view/object/?id=0e3a4810-514c-496c-afd3-7c65818ad373
https://publications-cnrc.canada.ca/fra/voir/objet/?id=0e3a4810-514c-496c-afd3-7c65818ad373

Access and use of this website and the material on it are subject to the Terms and Conditions set forth at https://nrc-publications.canada.ca/eng/copyright
READ THESE TERMS AND CONDITIONS CAREFULLY BEFORE USING THIS WEBSITE.

L'accès à ce site Web et l'utilisation de son contenu sont assujettis aux conditions présentées dans le site https://publications-cnrc.canada.ca/fra/droits
LISEZ CES CONDITIONS ATTENTIVEMENT AVANT D'UTILISER CE SITE WEB.

Questions? Contact the NRC Publications Archive team at PublicationsArchive-ArchivesPublications@nrc-cnrc.gc.ca. If you wish to email the authors directly, please see the first page of the publication for their contact information.

Vous avez des questions? Nous pouvons vous aider. Pour communiquer directement avec un auteur, consultez la première page de la revue dans laquelle son article a été publié afin de trouver ses coordonnées. Si vous n'arrivez pas à les repérer, communiquez avec nous à PublicationsArchive-ArchivesPublications@nrc-cnrc.gc.ca.

National Research Council Canada | Conseil national de recherches Canada

Canada

Kawennón:nis: the Wordmaker for Kanyen'kéha

Anna Kazantseva
National Research Council Canada
anna.kazantseva@nrc-cnrc.gc.ca

Owennatekha Brian Maracle
Onkwawenna Kentyohkwa
owennatekha@gmail.com

Ronkwe'tiyóhstha Josiah Maracle
Onkwawenna Kentyohkwa
diotsta@gmail.com

Aidan Pine
National Research Council Canada
aidan.pine@nrc-cnrc.gc.ca

Abstract

In this paper we describe preliminary work on *Kawennón:nis*, a verb conjugator for Kanyen'kéha (Ohsweken dialect). The project is the result of a collaboration between Onkwawenna Kentyohkwa Kanyen'kéha immersion school and the Canadian National Research Council's Indigenous Language Technology lab.

The purpose of Kawennón:nis is to build on the educational successes of the Onkwawenna Kentyohkwa school and develop a tool that assists students in learning how to conjugate verbs in Kanyen'kéha; a skill that is essential to mastering the language. Kawennón:nis is implemented with both web and mobile front-ends that communicate with an application programming interface that in turn communicates with a symbolic language model implemented as a finite state transducer. Eventually, it will serve as a foundation for several other applications for both Kanyen'kéha and other Iroquoian languages.

1 Introduction

Kanyen'kéha is an Iroquoian language, commonly known as "Mohawk", spoken in parts of Canada (Ontario and Quebec) and the United States (New York state). It has a vibrant community of learners, and educators but only about 3,500 L1 (first-language) speakers. Three main dialects are currently in use: Western (Ohsweken and Kenhteke), Eastern (Kahnawake, Kanehsatake and Wahta) and Central (Akwesasne). In our current work we focus exclusively on the Western dialect as spoken in Ohsweken.

The Truth and Reconciliation Committee of Canada led an inquiry into the atrocities committed during the residential school era in Canada from 1883 to the late 1990s. In 2015, they released a report (TRC, 2015) containing 94 calls to action, among them five action items that are related to Indigenous languages and culture. The report confirmed a macabre reality of residential schools that Indigenous people have known all along, that the residential school system was "created for the purpose of separating Aboriginal children from their families, in order to minimize and weaken family ties and cultural linkages" (TRC, 2015).

Since the release of the TRC report, a number of governmental and non-governmental programs and initiatives aimed at supporting Indigenous languages in Canada have been started. This project is a part of the National Research Council's (further *NRC*) related initiative to research and develop Indigenous language technology[1].

Despite the disproportionate duress that speakers of Indigenous languages have endured at the hands of Canadian colonial policies, the resilience and tenacity of Indigenous language communities can be seen in the myriad ways that they have resisted and continued to teach, learn and speak their languages (Pine and Turin, 2017, for further discussion). One such initiative is that of Onkwawenna Kentyohkwa, described in depth in Section 4. Given the widely celebrated accomplishments of Onkwawenna Kentyohkwa for almost two decades, it is apparent that it will be successful with or without the assistance of

This work is licensed under a Creative Commons Attribution 4.0 International License. License details: http://creativecommons.org/licenses/by/4.0/

[1]For more information about the project see www.nrc-cnrc.gc.ca/eng/solutions/collaborative/indigenous_languages/index.html

Proceedings of Workshop on Polysynthetic Languages, pages 53–64
Santa Fe, New Mexico, USA, August 20-26, 2018.

technology. It is therefore the goal of Kawennón:nis to develop a fundamentally assistive reference tool that doesn't *replace* learners' experience of acquiring Kanyen'kéha verbal morphology at Onkwawenna Kentyohkwa, but that *augments* and *complements* it.

Another factor contributing to the technological inequity faced by second-language learners of Kanyen'kéha compared with learners of more resourced languages is the result of the text-based bias of contemporary language technologies. Similar to most other Indigenous languages of North America, Kanyen'kéha acquired a standardized orthography recently (1993), relative to the multi-millennial history of the language's oral tradition. Dialectal differences, variations in spelling and the general paucity of written content in the language make statistical language learning difficult. The situation is particularly dire for the Western dialect: very little text is available in the Ohsweken dialect of Kanyen'kéha apart from Onkwawenna Kentyohkwa's curricular materials.

Most Indigenous languages in Eastern, Central and Northern Canada show high levels of polysynthesis, that is, their words are composed of relatively many morphemes. All Iroquoian languages fall into this category. Single words in Kanyen'kéha are routinely translated as entire sentences in English. For example:

(1) *enhake'serehtsherakwatákwahse-'*
en-hake-'serehtsher-kwatakw-ahse-'
will-he.to.me-car-fix.it.up-for.me-punctual
FUT-3SG.M/1SG-fix.it.up-BEN-PUNC

'he will repair my car for me'

(2) *tetsyonkyathahahkwahnónhne*
te-ts-yonky-at-hahahkw-hnón-hne
both-again-it.to.you.and.I-each.other-walk-purposive-have.gone
DUAL-REP-3SG.N/1.DU.INCL-SREFL-walk-PURP-PPFV

'the two of us went for a walk'

The generative power and compositional morphology of the language is such that neologisms are easily created, with no borrowing necessary. For example, the word for *pizza* is *teyotena'tarahwe'nón:ni* which literally means *round bread*. However, this property also makes the language difficult to teach, as learners cannot be expected to memorize paradigms arbitrarily. Rather, they must learn the patterns to *generate* the paradigms. This is the essential insight on which the morpheme-based teaching approach adopted by the Onkwawenna Kentyohkwa is based (see Section 4). The fact that there are 72 distinct bound pronominal morphemes in Kanyen'kéha means that even learning how to properly conjugate a modest number of verbs requires a significant amount of work. Kawennón:nis is a tool designed to help students with this task. It allows a user to enter a number of variables which the software then processes and outputs the corresponding grammatical form.

To the best of our knowledge, textual data in the Ohsweken dialect of Kanyen'kéha available publicly is quite scarce. Accordingly, we were not able to identify a sufficiently large corpus to use with statistical methods. The Ohsweken dialect also has less linguistic documentation than the Eastern and Central dialects which have been studied more extensively (Postal, 1963; Bonvillain, 1973; Beatly, 1974; Mithun, 2004). However, Onkwawenna Kentyohkwa has created a textbook (Maracle, 2017) that describes and explains salient morphological properties of Kanyen'kéha as spoken in Ohsweken. The textbook, along with linguistic sources about other dialects of Kanyen'kéha and the closely related language Oneida are the primary sources of information for creating the symbolic language model of Ohsweken Kanyen'kéha.

We see the main contributions of this work as follows. Successful collaborations between Indigenous communities and technology partners are rare. Perhaps the single most important contribution of this work is that it gives an example of a productive, harmonious collaboration between an immersion school

within an Indigenous community (Onkwawenna Kentyohkwa) and a research group from outside the community (NRC). Details of this collaboration are discussed in Section 4.

Secondly, Kanyen'kéha has very little software support. Apart from LanguageGeek[2] and FirstVoices[3] keyboards there is no software that enables use of the language on desktops or on mobile devices. It is therefore a second important contribution of this work that we are creating the first computational language model of Kanyen'kéha.

The third contribution is in the applications, though at present these are prototypes and proofs of concept. Creating a symbolic language model is time consuming and error-prone. The finite state transducer behind Kawennón:nis is currently used for the verb conjugator and a limited spell-checker but in future, we hope to add other applications, thereby maximally leveraging the time and effort involved in making Kawennón:nis.

This paper is structured as follows. In Section 2 we place our project in the context of related research. In Section 3 we provide a high-level sketch of the relevant Kanyen'kéha verbal paradigms. Section 4 describes the immersion school Onkwawenna Kentyohkwa, the unique immersion teaching method it uses and the collaboration with NRC. Section 5 provides an overview of the software architecture of Kawennón:nis. In Section 6 we describe the symbolic language model that we have created. Section 7 talks about two prototype applications implemented to date and includes a brief discussion of evaluation. Finally, we discussing future Work in Section 8 and provide conclusions in Section 9.

2 Related Work

In this section we will list some of the relevant computational systems. Our work also heavily depends on linguistic sources (Lounsbury, 1953; Postal, 1963; Bonvillain, 1973; Beatly, 1974; Mithun, 1996; Mithun, 2004)). Furthermore, many language activists, teachers, students have made important contributions which, for the most part, are not available for citing.

While in general it is true that Indigenous languages lack adequate linguistic software support, our work relies on previous computational and linguistic research.

Moshagen et al. (2013) describe Giella - a framework for FST-based modeling of languages with the aim of easily creating end-user applications. The Giella infrastructure has been successfully used to create FSTs and corresponding tools for a number of languages: Cree (Snoek et al., 2014; Harrigan et al., 2017), Northern Haida (Lachler et al., 2018), Odawa (Bowers et al., 2017) are but a few of the many examples.

In a different line of work Littell et al. (2017) have created a tool called *Mother Tongues Dictionaries* (formerly *Waldayu*) that helps communities create web and mobile dictionaries from potentially heterogeneous community resources. The mobile version is used as the mobile front end for the extensive FirstVoices resources (Brand et al., 2015) .

For Indigenous languages where it is realistic to collect sufficient data, statistical modeling has been used successfully. Martin et al. (2003; Désilets et al. (2008) use parallel English-Inuktitut corpora to create translation memories and Micher (2017) uses a monolingual corpus of Inuktitut to improve the performance of a morphological analyzer.

For a recent survey of language technologies for Indigenous languages in Canada see (Littell et al., forthcoming). A more extensive inventory of open-source resources, in both Indigenous and other languages, is available at `github.com/RichardLitt/endangered-languages`.

Our work is most similar to that of Snoek et al. (2014; Harrigan et al. (2017; Lachler et al. (2018; Bowers et al. (2017). Kawennón:nis is somewhat similar to morphologically-aware dictionaries, but its design and content is far more customized.

[2] `http://www.languagegeek.com/rotinonhsonni/keyboards/iro_keyboards.html`
[3] http://www.firstvoices.com/en/apps

Figure 1: Examples of bound pronouns in Kanyen'kéha

Bound pronoun in Kanyen'kéha	*English translation*	*Example (Kanyen'kéha)*	*Example (English)*
ke-	me to it	**ke**khón:nis	I cook
ra-	he to it	**ra**khón:nis	he cooks
yonkeni-	it to me and you (or smb.)	**yonkeni**nòn:we's	it likes us (both)
yako-	it to her	**yako**nòn:we's	it likes her
take-	you to me	**take**nòn:we's	you like me
she-	you to her	**she**nòn:we's	you like her

3 Kanyen'kéha, the language of the Kanyen'kehá:ka

Kanyen'kéha[4] is part of the Iroquoian language family[5]. Two main subgroups of the Iroquoian language family are distinguished: the Southern Iroquoian (now only containing Cherokee) and Northern Iroquoian (Cayuga, Onondaga, Seneca, Mohawk (Kanyen'kéha), Oneida, Wyandot, Nottoway and Tuscarora) (Mithun, 2004). Wyandot is currently often considered a sleeping language, but see (Lukaniec, 2010) for very promising reclamation and revitalization work.

Kanyen'kéha has 11 consonants[6] /h/, /k/, /n/, /r/, /s/, /t/, /w/, /y/, /ts/, /k^{w}/, and /ʔ/ (Mithun, 2004) and 6 vowels /a/, /e/, /i/, /o/, /ʌ̃/, and /õ/[7]. Vowels can be marked for stress (ˊ) and falling tone (ˋ); both stress and falling tone can be marked for length (*:*) (Mithun, 1996).

A pronominal prefix, a verb root and an aspectual ending are always present (for commands the aspectual ending is null). A verb can contain pre- and post-pronominal prefixes and pre-aspectual suffixes. Verb roots are also bound morphemes and do not, on their own, constitute well-formed words.

Kanyen'kéha has 15 stand-alone or *free pronouns*, and 72 *bound pronouns* which can only be used as a part of a verb. Bound pronouns in Kanyen'kéha are very complex (Mithun, 1996). In one morphological unit, a pronoun encodes information about both the agent and the patient of an action. For both the agent and the patient, it also captures gender (male, female, neuter), number (single, dual or plural) and whether the hearer is included in the set (inclusivity). If no agent or patient is available, the pronoun is the same as the single neuter pronoun.

Some example pronouns are shown in Figure 1. The bound pronouns are divided into three groups: active pronouns (3; roughly for situations where the actor is human, and the patient is not, or where there is no patient), passive pronouns (4; for situations where the patient is human and the agent is not) and "transitive" pronouns (5; for situations when both the actor and the patient are human). Below are some examples:

(3) ***Se**nòn:wes*
se-nonhwe'n-s
you.to.it-like-habitual
2.SG.AGENT-like-HAB

'**You** like it.'

(4) ***Sa**nòn:wes*
sa-nonhwe'n-s
it.to.you-like-habitual
2.SG.PATIENT-like-HAB

'**It** likes you.'

[4]ISO 639-2 code is *moh*.

[5]In this Section and further in this paper we are sometimes faced with the choice of whether to use standardized terminology from Linguistics, or to rely on the knowledge of the community. Almost exclusively, we use the terminology and categories as described in (Maracle, 2017). We made this decision because the textbook is the most detailed and accurate description of the Ohsweken dialect available to-date and because the theoretical understandings of the textbook form the basis of our language model.

[6]These are represented orthographically and taught as 8 consonants, h, k, n, r, s, t, w, y, plus glottal stop ' (Maracle, 2017).

[7]These are represented orthographically as a, e, i, o, en, on respectively.

(5) ***Take**nòn:wes*
take-nonhwe'n-s
you.to.me-like-habitual
2.SG/1.SG-like-HAB

'**You** like **me**.'

In reality, the situation is not quite this simple: in many cases, the choice of a pronoun is lexicalized and determined more by the history of a specific verb or by traditional usage than by tangible properties of the actor or the patient.

For instance the verb *rihwayent* meaning 'decide' as seen in example 6 below uses patientive pronouns (in this case *wake*, roughly meaning *it to me*.)

(6) *ya'tewakerihwayentà:ses*
y-a'-te-wake-rihwayent-'se-s
there-did-it.to.me-decide-for.me-habitual
TRANS-FACT-1.SG.PATIENT-business.matters-decide-BEN-HAB

'I decided.'

There are 11 possible pre-pronominal prefixes and a verb can take on one or more of them. Some possible prefixes are the definite (7), cislocative (8), translocative (9), and many others. For example:

(7) *Wa'katáweya'te'*
wa'-k-ataweya't-e'
did-I.to.it-enter.a.place-punctual
FACT-1.SG.AGENT-enter.a.place-PUNC

'I entered (a place).'

(8) ***T**akatáweya'te'*
t-a-k-ataweya't-e'
here-did.here-I.to.it-enter.a.place-punctual
CIS-FACT-1.SG.AGENT-enter.a.place-PUNC

'I came in (**here**).'

(9) ***Y**a'katáweya'te'*
y-a'-k-ataweya't-e'
there-did-I.to.it-enter.a.place-punctual
TRANS-FACT-1.SG.AGENT-enter.a.place-PUNC

'I went in (**there**).'

A verb may also take one of the three pre-aspectual case suffixes (Beatly, 1974). For example the distributive (11), reversive (12) or purposive (13):

(10) *Enkhnyó:ten'*
en-k-hnyot-en'
will-I-stand.up-punctual
FUT-1SG.AGENT-stand.up-PUNC

'I will stand it up.'

(11) *Enkhny**ónnyon**'*
en-k-hnyot-**onnyon**-'
will-I-stand.up-**distributive**-punctual
FUT-1SG.AGENT-stand.up-**DIST**-PUNC

'I will stand **many things** up.'

(12) *Enkhnyot**á:ko**'*
en-k-hnyot-**ako**-'
will-I-stand.up-**reversive**-punctual
FUT-1SG.AGENT-stand.up-**REV**-PUNC

'I will **lower** it.'

(13) *Enkhnyot**à:n**a'*
en-k-hnyot-**ahn**-a'
will-I-stand.up-**purposive**-punctual
FUT-1SG.AGENT-stand.up-**PURP**-PUNC

'I will **go** stand it up.'

The stem of a verb can also be complex due to noun incorporation. In the current version of Kawennón:nis we only cover a limited set of paradigms: namely the mandatory parts of the verb (the pronominal prefix, the root and the aspectual ending). In the coming months, we will extend our model to all paradigms that do not require semantic restrictions.

4 Collaboration between NRC & Onkwawenna Kentyohkwa

Onkwawenna Kentyohkwa is an adult immersion program in southern Ontario that was founded in 1999. The program currently teaches a two-year program that enables students to achieve high levels of proficiency in Kanyen'kéha (the "Mohawk" language). Students attend six hours per day, five days a week, for two school years, totaling 2,000 hours of instruction, all in the language.

The program uses a unique morpheme-based teaching method that allows students with little or no previous language exposure to achieve one of the Advanced levels on oral proficiency assessments based on the American Council on Teaching of Foreign Languages model (Breiner-Sanders et al., 2000). A few graduates are now raising children as first-language speakers.

The program's success has generated interest across the continent and adult groups from other Iroquois communities (Oneida and Seneca) have translated the curriculum and materials into their languages to teach adults in their communities.

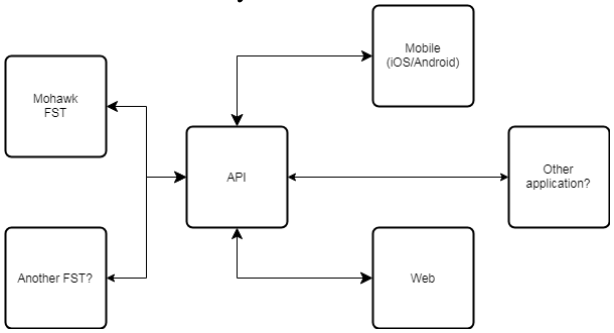
The idea behind creating Kawennón:nis, originated with a staff member of the Onkwawenna Kentyohkwa School. Thus, the Onkwawenna Kentyohkwa school defined the scope and the main parts of Kawennón:nis, motivated by a specific educational need that the staff of Onkwawenna Kentyohkwa identified as important.

In order to move ahead with the project in a collaborative fashion, NRC partnered with a staff member of the school who participated in ongoing correspondence and communication, including weekly online meetings to discuss the user interface of Kawennón:nis and to make decisions regarding morphology. Additionally, members of NRC have made several in-person visits to Ohsweken to demonstrate Kawennón:nis to students and staff of Onkwawenna Kentyohkwa and to participate in intensive, multi-day, collaborative efforts brainstorming project decisions and direction as well as designing the user interfaces. This level of mutual collaboration and continual involvement from both NRC and Onkwawenna Kentyohkwa will be essential if the project is to succeed.

5 Kawennón:nis: Overview

Kawennón:nis is built with the hope of supporting a wide variety of user applications, and of eventually extending it to use with other Iroquoian languages. It consists of three main parts: (1) a 'frontend' implemented with Angular and Ionic for the web, as well as Android and iOS platforms. (2) An API implemented in Python (Flask). (3) a 'backend' consisting of a finite state transducer implemented in Foma (Hulden, 2009). This architecture allows alternate frontends or future iterations of the current frontend tools to make use of the FST. It also gives a clear point of integration for FSTs of other Iroquoian languages. This section describes the software architecture for Kawennón:nis as illustrated in Figure 2.

Figure 2: Kawennón:nis system architecture from back to front.



5.1 API

The Kawennón:nis API (Application Programming Interface) is used to define the methods through which frontend components or applications like websites or mobile apps may interact with the backend FST. The API is read-only, requires an API-key to be accessed and follows RESTful design principles, whereby the API is stateless, cacheable, decoupled from the FST and frontend applications as much as possible and exposes information about available verbs, pronouns, conjugated forms and other data as discrete resources and subresources.

5.2 Web & mobile applications

The user interface for both web and mobile frontends have the same core components. First, a 'conjugation' component which renders the API response on a number of different user-defined tiers. Second, a 'palette' component which allows the user to set the parameters for the request to the API. Below in Figure 3 the conjugation component can be seen on the left and the palette component on the right.

Additionally, the user has control over which parts in the conjugation to highlight. The Onkwawenna Kentyohkwa immersion school uses a colour scheme where agents are highlighted in red, patients in blue and transitive pronouns are highlighted in purple. For ensuring an easy transition between the course and Kawennón:nis, this colour scheme is preserved in Kawennón:nis and the user can optionally select whether to also colour-code the verb stem or various other types of affixes (aspectual suffixes, for example). The user is also able to select between a variety of tiers to display in the conjugation component. The first, main tier is the orthographic word without any segmentation. The next tiers that are available include a morpheme breakdown tier, and two type glossing tiers, one which simply glosses all morphemes by their category (ie 'pronoun', 'root' etc.) and another which gives the English gloss (ie. 'you', 'cook' etc.). Users are anonymously authenticated and their settings are saved across sessions on the same devices.

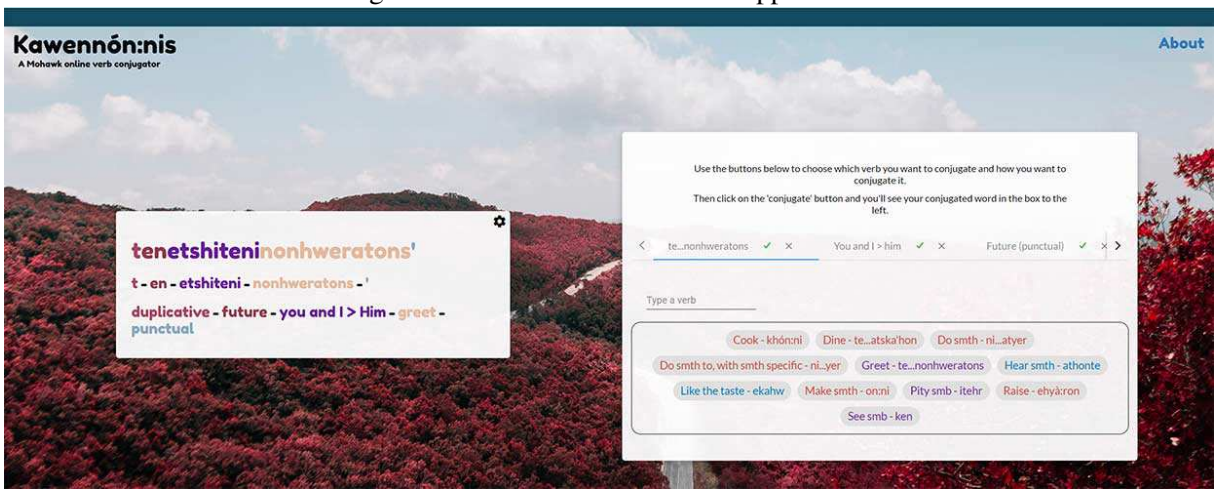
6 Symbolic Model of Verbal Morphology in Kanyen'kéha

The backbone of Kawennón:nis is a symbolic model of the Kanyen'kéha verbal morphology implemented as a *finite state transducer* (further *FST*).

A finite state transducer is a type of finite state automaton that maps between two sets of states: the input set and the output set. The sets of states are often referred to as the input and output alphabets. The transducer can be thought of as a "translating machine": it translates an input sequence into an output one. In our context, the FST first translates a sequence specifying 1) the verb stem and 2) a set of tags capturing the desired morphological properties, and then it outputs a verb correct form.

Formally, an FST is a seven-tuple $(Q, \Sigma, \Gamma, \delta, w, I, F)$ such that Q is the finite set of all possible states, Σ is the finite set of all possible input symbols – the input alphabet, Γ is the finite set of all possible output

Figure 3: The Kawennón:nis web application



symbols – the output alphabet, $\delta : Q \times \Sigma \rightarrow Q$ is the transition function, $w : Q \times \Sigma \rightarrow \Gamma$ is the output function, $I \subset Q$ is the set of initial states and $F \subset Q$ is the set of accepting states (Mohri, 1997).

For example, a toy FST for a subset of conjugations of the verb *cook* is shown in Figure 4 below. Example 14 below is a demonstration of the valid input mappings to the corresponding output sequences.

(14) *a*. cook+3rd → cooks

b. cook+Past → cooked

c. kick+Participle → kicking

In general even if the morphological paradigms that are encoded are complex, adding new verbs is simple and straightforward.

Following the convention (Beesley and Karttunen, 2003; Koskenniemi, 1986), we separate lexical and morphological rules from phonological alternations in our implementation. The latter are applied as a second layer. The rules are implemented as a continuation lexicon in *lexc* formalism and the phonological alternations as rewrite rules, both of which are implemented in Foma (Hulden, 2009). See Beesley and Karttunen (2003) for a discussion on structuring linguistic FSTs.

Creating a linguistic FST involves manually hand-coding all morphological rules - a very time-consuming and potentially error-prone process. A more contemporary alternative would be to learn morphology from a corpus. However, we are unaware of any sufficiently large, homogeneous corpus of Kanyen'kéha; nor are we aware of any corpus at all for the Ohsweken dialect. On the other hand, some linguistic documentation for Kanyen'kéha is available (Beatly, 1974; Bonvillain, 1973; Postal, 1963; Mithun, 1996; Mithun, 2004). The root-word method textbook (Maracle, 2017), the linguistic sources listed above and continuous input from a proficient speaker of Kanyen'kéha[8] have been our main sources of information.

The FST is structured so that adding new verbs is fast and painless: it amounts to entering a lexical entry into a spreadsheet and specifying several properties. The input, collation and quality control of this task is done by both NRC researchers and teachers at Onkwawenna Kentyohkwa.

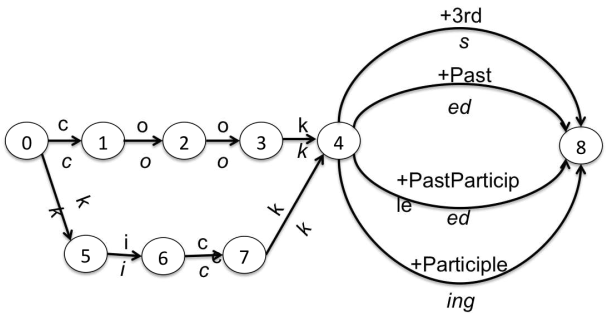
Long-distance dependencies complicate the otherwise relatively straightforward structure of morphological rules. While there are no true long-distance dependencies, the selectional preferences of verbs are treated *like* long-distance dependencies by the *lexc* formalism. Some examples are restrictions on what type of bound pronoun can be used with a given verb, the conjugation class of a verb or the presence of a mandatory prefix. We model such dependencies using flag diacritics.

Phonological alternations are yet another challenge. For example, each of the 72 possible bound pronouns changes depending on which sound follows it.[9] Five major categories can be distinguished (verb stems that start with a consonant, an "a", "e', "i", or an "o'). However, there are numerous exceptions to this which makes encoding phonological rules non-trivial.

[8]The 3rd author of this paper.

[9]In the current version of Kawennón:nis this placeholder can only be taken by the verb root. However, in reality, other affixes are possible.

Figure 4: A toy finite-state transducer. A sequence *cook+Past* is translated as *cooked*.



Kawennón:nis is an early-stage research project and only a limited prototype has been released (to a limited audience) and demonstrated to the community. The current version contains 30 verb stems, all bound pronouns and 6 temporal paradigms (Command, Habitual form, Perfective form and Punctual forms which include Definite past, Conditional and Future forms). These were chosen because of their direct and early relevance in the teaching curriculum at Onkwawenna Kentyohkwa. We are working on extending the lexicon to approximately 500 verb roots and including other paradigms.

7 Applications and Evaluation of the Kawennón:nis FST

Because of the early stage of this work and the absence of corpora, Kawennón:nis has not yet been formally evaluated. However, our current workflow involves weekly meetings and consultations with a proficient language speaker, who is a recent graduate of the program and now an educator[10]. The correctness of the rules and the generated forms is checked step-by-step, manually. To the extent possible, we intend to run a statistical evaluation once Kawennón:nis is close to completion.

Currently, the FST model is the backbone of three applications: the web application Kawennón:nis, its mobile app version and a spellchecker. We have already described Kawennón:nis in Section 5. The spellchecker is generated using the Giellatekno infrastructure (Moshagen et al., 2013). The spellchecker only works in Office Libre but it is also unclear if the spellchecker will end up being released or used. There is some hesitance within the community to develop a tool that corrects users' spelling, as there is this approach might be too prescriptive and punitive as well as reinforce a singular dialect, not allowing for minor, but accepted, idiosyncratic differences in spelling. That is to say, just because we *can* develop a particular technology with little extra effort does not necessarily mean we *should*. This question requires further consideration.

Given how time-consuming it is to create a symbolic language model, we intend to use it for as many applications as possible. In the near future, we intend to start a flash-card generator that would allow students learn the subsets of morphology they are most interested in.

The next application we intend to focus on is predictive text on mobile devices.

8 Future Work

In the immediate future we will extend the inventory of verb roots to 500 and work on including all relevant paradigms, namely pre-pronominal prefixes and pre-aspectual suffixes.

It is quite likely that we will be unable to include a majority of derivational affixes because selectional preferences are not uniform across verbs in Kanyen'kéha. This may be addressed to a large extent by subcategorizing the lexicon but this solution may become too unwieldy. In a similar vein, we will probably be unable to properly address noun incorporation due to semantic restrictions. However, we hope to include a small subset of derivational affixes in a held-out, limited set of verbs that will help students master the phenomena of derivational suffixes and noun incorporation.

Although nouns are not as frequent in Kanyen'kéha as verbs, they are obviously still a very important part of the language and thus we will encode some nominal morphology after the verbal part is complete.

One of the biggest problems in collecting a reliable corpus of Kanyen'kéha is dialectal differences. Plans for future work include creating a transducer between dialects which would allow corpora from different dialects to be leveraged and used for statistical learning. It is also our hope that this work be extended to other Iroquoian languages, such as Oneida.

While most of the future work mentioned above is centered around improving the language model, we are also interested in leveraging the generative power of the model by making use of it in other applications. The details will not be clear until Kawennón:nis has been launched and used by the Onkwawenna Kentyohkwa immersion school. That is, what was needed in the initial applications of Kawennón:nis was ascertained by Onkwawenna Kentyohkwa staff who have the perspective that only years of experience can provide, and similarly, future applications and implementations of Kawennón:nis following its initial release will require thoughtful planning and an acute awareness of the needs of Kanyen'kéha teach-

[10]This person is the third author of this paper

ers and learners alike. Some potential applications include automatic flash card generators, spellcheckers and predictive text for mobile devices.

9 Conclusions

In this paper we presented preliminary work on Kawennón:nis, a verb conjugator for Kanyen'kéha.

The FST technology behind Kawennón:nis is time-consuming to engineer and difficult to change or adapt, yet the low-data reality of many Indigenous languages is such that we cannot rely on statistical alternatives. In general, smaller languages do not benefit in the short term from recent advances in Language Processing or Machine Learning, due to a general scarcity of linguistic data. Restricting our toolset to newer technologies would preclude us from making progress on this project.

Looking beyond Kawennón:nis, it is important to note the general lack of linguistic software support for most Indigenous languages of North America, with many language communities only recently having their writing systems supported by Unicode or having an input system to write their languages on mobile devices[11] or computers (Pine and Turin, 2018).

With Indigenous languages facing increased pressure from majority languages like English or French, many language activists are turning to technology to assist in reclaiming and revitalizing their languages. Developing basic technological scaffolding to support online communication through social media, text processing tools and reference tools like Kawennón:nis could have an important role to play in this process. It is clear that given the self-determined nature of Indigenous language reclamation efforts that the only way forward is through respectful, contextually-informed collaboration. It is our hope that the development of Kawennón:nis is a small but confident step in the right direction.

Acknowledgements

The authors are grateful Karakwenhawi Zoe Hopkins for allowing the first author to participate in the Kanyen'kéha online course. Many thanks to Antti Arppe and Sjur Moshagen for sharing their experience and for advice on FSTs, and to Jordan Lachler for extensive help on the Linguistics of Iroquoian languages. We are also grateful to Rohahí:yo Jordan Brant and to Ryan DeCaire for patient help with Kanyen'kéha and for multiple suggestions regarding the design of the user interface.

Glossary

CIS - cislocative, DUAL - dualic, FACT - factual, FUT - future, PPFV - past perfective, PUNC - punctual, PURP - purposive, REP - repetitive, REV - reversive, SREFL - semi-reflexive, TRANS - translocative.

References

John Beatly. 1974. *Mohawk Morphology*. Number 2 in Linguistic Series. Museum of Anthropology, University of Northern Colorado, Greeley, Colorado.

Kenneth R. Beesley and Lauri Karttunen. 2003. *Finite State Morphology*. CSLI Publications.

Nancy Bonvillain. 1973. *A Grammar of Akwesasne Mohawk*. Number 8 in Ethnology Division. National Museum of Man, Ottawa, Canada.

Dustin Bowers, Antti Arppe, Jordan Lachler, Sjur Moshagen, and Trond Trosterud. 2017. A morphological parser for Odawa. In *Proceedings of 2nd Workshop on Computational Methods for Endangered Languages (ComputEL-2)*.

Peter Brand, Tracey Herbert, and Shay Boechler. 2015. Language vitalization through online and mobile technologies in british columbia. In Laurel Evelyn Dyson, Stephen Grant, and Max Hendriks, editors, *Indigenous people and mobile technologies*, chapter 17. Routledge.

Karen E. Breiner-Sanders, Pardee Lowe, John Miles, and Elvira Swender. 2000. Actfl proficiency guidelinesspeaking: Revised 1999. *Foreign Language Annals*, 33(1):13–18.

[11] With FirstVoices Keyboards and Keyman.

Alain Désilets, Benoit Farley, Geneviève Patenaude, and Marta Stojanovic. 2008. WeBiText: Building large heterogeneous translation memories from parallel web content. *Proc. of Translating and the Computer*, 30:27–28.

Atticus G. Harrigan, Katherine Schmirler, Antti Arppe, Lene Antonsen, Trond Trosterud, and Arok Wolvengrey. 2017. Learning from the computational modelling of Plains Cree verbs. *Morphology*, 27(4):565–598.

Mans Hulden. 2009. Foma: a finite-state compiler and library. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 29–32. Association for Computational Linguistics.

Barbara Kelly, Gillian Wigglesworth, Rachel Nordlinger, and Joseph Blythe. 2014. The acquisition of polysynthetic languages. *Language and Linguistics Compass*, 8(2):51–64.

Kimmo Koskenniemi. 1986. Compilation of automata from morphological two-level rules. In F. Karlson, editor, *Papers from the Fifth Scandinavian Conference on Computational Linguistics*, pages 143–149.

Jordan Lachler, Lene Antonsen, Trond Trosterud, Sjur N. Moshagen, and Antti Arppe. 2018. Modeling Northern Haida morphology. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2018)*, May.

Patrick Littell, Aidan Pine, and Henry Davis. 2017. Waldayu and Waldayu Mobile: Modern digital dictionary interfaces for endangered languages. In *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 141–150.

Patrick Littell, Anna Kazantseva, Roland Kuhn, Aidan Pine, Antti Arppe, Christopher Cox, and Marie-Odile Junker. forthcoming. Indigenous language technologies in Canada: Assessment, challenges, and successes. *Proceedings from the 27th International Conference on Computational Linguistics*.

Floyd Lounsbury. 1953. *Oneida Verb Morphology*. Yale University Press.

Megan Lukaniec. 2010. Words of the Huron. By John L. Steckley. *International Journal of American Linguistics*, 76(2):304–306.

Brian Maracle. 2017. *Anonymous 1st Year Adult Immersion Program 2017-18*. Onkwawenna Kentyohkwa, Ohsweken, ON, Canada. The book was co-written my several other staff members over the years. Brian Maracle is the author of the latest, 2017 edition.

Joel Martin, Howard Johnson, Benoit Farley, and Anna Maclachlan. 2003. Aligning and using an English-Inuktitut parallel corpus. In *Proceedings of the HLT-NAACL 2003 Workshop on Building and using parallel texts: Data driven machine translation and beyond, Volume 3*, pages 115–118. Association for Computational Linguistics.

Jeffrey Micher. 2017. Improving coverage of an Inuktitut morphological analyzer using a segmental recurrent neural network. In *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 101–106. Association for Computational Linguistics.

Marianne Mithun. 1989. The acquisition of polysynthesis. *Journal of Child Language*, 16(2):285–312.

Marianne Mithun. 1996. Grammatical Sketches: the Mohawk Language. In *Quebec's Aboriginal Languages*, pages 159–174. Multilingual Matters Ltd.

Marianne Mithun. 2004. Mohawk and the iroquoian languages. In *Routledge Encyclopedia of Linguistics*. New York: Routledge.

Mehryar Mohri. 1997. Finite-state transducers in language and speech processing. *Computational Linguistics*, 23(2):269–311, June.

Sjur N. Moshagen, Tommi A. Pirinen, and Trond Trosterud. 2013. Building an open-source development infrastructure for language technology projects. In *Proceedings of the 19th Nordic Conference of Computational Linguistics, NODALIDA 2013, May 22-24, 2013, Oslo University, Norway*, pages 343–352.

Aidan Pine and Mark Turin. 2017. Language revitalization. *Oxford Research Encyclopedia of Linguistics*.

Aidan Pine and Mark Turin. 2018. Seeing the heiltsuk orthography from font encoding through to unicode: A case study using convertextract. *3rd Workshop on Collaboration and Computing for Under-Resourced Languages 'Sustaining knowledge diversity in the digital age'*.

Paul M. Postal. 1963. *Some Syntactic Rules in Mohawk*. Ph.D. thesis, Yale University.

Conor Snoek, Dorothy Thunder, Kaidi Lõo, Antti Arppe, Jordan Lachler, Sjur Moshagen, and Trond Trosterud. 2014. Modeling the noun morphology of Plains Cree. In *Proceedings of the 2014 Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 34–42, Baltimore, Maryland, USA, June. Association for Computational Linguistics.

TRC. 2015. Truth and Reconciliation Commission of Canada: executive summary. Truth and Reconciliation Commission of Canada, Manitoba: Truth and Reconciliation Commission of Canada.