Empowering Oneida Language Revitalization: Development of An Oneida Verb Conjugator

Yanfei Lu[1], Patrick Littell[2], Keren Rice[1]
University of Toronto[1], National Research Council Canada[2]
100 St George St., Toronto, ON, M5S 3G3, 1200 Montreal Road, Ottawa ON, K1A 0R6
yanfei.lu@mail.utoronto.ca, patrick.littell@nrc-cnrc.gc.ca, rice@chass.utoronto.ca

Abstract

In this paper, we present the development of a digital Oneida verb conjugator through using the Gramble framework. This project is a collaborative effort with the Twatati Adult Oneida Language program. Oneida is a polysynthetic North American Indigenous language. Its verb roots can be conjugated with multiple affixes, and long verbal complexes can be used as utterances. Each Oneida affix encodes important grammatical information, and its form often varies based on various factors, such as its position in the utterance and its phonological environment. The distinct morphosyntactic structures complicate acquisition of the language by learners who are native speakers of English. With an alarmingly small number of native speakers of Oneida, supporting and accelerating adult second language leaners' acquisition process has become a pressing necessity. The Oneida verb conjugator can demonstrate its users the correct conjugations of verbs and can also let learners generate practice materials tailored to their unique learning trajectories. This paper presents the preliminary stages and outcomes of the project and outlines the areas for improvement to be addressed in our subsequent endeavors.

Keywords: Indigenous language revitalization, Verb conjugator, Polysynthetic language

1. Introduction

Oneida is the language of the Oneida people of North America. The Oneida people live mainly in three communities at present day: Oneida Nation of the Thames (near London, Ontario) and two Oneida reservations, one in Wisconsin and the other in New York (Michelson and Doxtator, 2002, p. 1). Oneida currently has a very small number of native speakers: only 45 in Canada (Statistics Canada, 2022) and 102 worldwide (Eberhard et al., 2023). According to elders from the Oneida Nation of the Thames, the number of fluent native speakers might actually be less than 20 today, while the youngest among them is in their 60s (Nancy George, p.c.). The *UNESCO Atlas of the World's Languages in Danger* has marked Oneida as critically endangered (UNESCO, 2009). To revitalize the language and continue the intergenerational transmission of Oneida, it is essential to help adult learners become fluent second language (L2) speakers and support them to pass the language to the younger generation in their family and community. This is because "[i]f we want our children to speak the language, those who shape them (adults) need to speak it" (DeCaire, 2023, p. 17).

With the goal of contributing to the revitalization of the Oneida language through developing language learning tools dedicated for adult L2 Oneida learners, we are collaborating with the Twatati Adult Oneida Language program to build a digital Oneida verb conjugator. Upon completion of the project, full ownership will go to the Oneida community, and it will be freely accessible to the public. Compared to physical resources such as textbooks and dictionaries, digital tools in the format of websites and mobile applications that can parse or generate the language automatically presents many advantages, which we will discuss in detail later. After a few simple clicks, a conjugator can demonstrate to its users how to conjugate verb roots with the correct affixes instantly. Learners could also use the verb conjugator as a tool for generating practice materials tailored to their own learning processes.

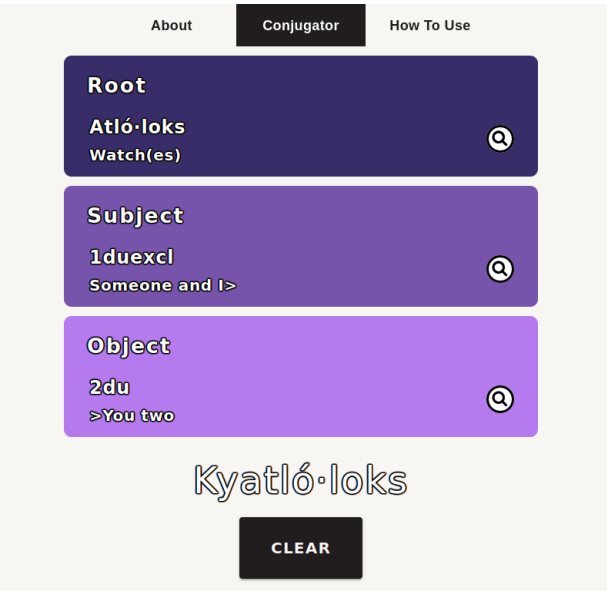

Figure 1: User interface of the Oneida verb conjugator (pilot version)

2. Background

2.1 Language Overview

Oneida is a member of the Iroquoian language family, and it is most closely related to Kanyen'kéha (Mohawk). Like all other languages of the family, Oneida is a polysynthetic language. This type of language uses complex morphemes to express meanings that would be expressed by individual words and sentences in languages like English, French, or Mandarin (O'Grady and Archibald, 2016, p. 318). As illustrated by example (1), the equivalent meaning of the English sentence 'Don't they all know

LREC-COLING 2024, pages 5757–5767

what they call (it)?' is expressed by a series of particles, verb roots, and affixes in Oneida[1].

(1) Yah kʌ tehonanúhte' oh náhte' kuwa·yáts?

yah	kʌ	te-	hon-	-anuhte-'
NEG	Q	PREPPR	3SGN>3PLM-	know-HAB
oh	nahte'	kuwa-		-yat-s
PTCL	what	3PLFI>3SGFZ		call-STAT

'Don't they all know what they call (it)?'

(adapted from Twatati, 2017)

Comparing this Oneida utterance[2] with its English counterpart, we can easily detect the significant differences between the structures of the two languages. These profound structural distinctions can lead to considerable challenges for adult native English speakers who embark on the journey towards mastery of the Oneida language. The small population of native speakers and fluent L2 speakers also intensifies the challenges to the L2 leaners since they have very limited opportunities to practice Oneida outside of the classroom. These issues can hinder learners' progress towards achieving advanced competency.

2.2 Phonemic Inventory and Orthography

Oneida has six vowels (orthographic representations are shown in brackets < >): /i/ <i>, /e/ <e>, /ɑ̃/ <ʌ>, /a/ <a>, /ũ/ <u>, /o/ <o>; and nine consonants: /t/ <t>, /n/ <n>, /s/ <s>, /l/ <l>, /j/ <y>, /k/ <k>, /w/ <w>, /ʔ/ <'>, and h <h> (Lounsbury, 1976; Julian, 2010; Michelson, 1983)[3]. In terms of the suprasegmental features, vowel lengthening is marked by the interpunct symbol <·>, which goes immediately after the lengthened vowel; stress is marked by the acute stress symbol <ˊ>, which goes on top of the stressed vowel; Devoicing is marked by underlining the devoiced segments (Lounsbury, 1976; Michelson, 1983; Julian, 2010; Michelson et al., 2016).

The Oneida language was originally passed down throughout the generations without a standard orthography. Several different orthography systems have been developed based on English during the past centuries since early contact between the Haudenosaunee people and the Europeans (Abbott and Metoxen, 2012, p. 3). The orthographies and diacritics used in different literature and documentations contain many variations (Abbott, 2016, p. 170; Michelson et al., 2016, p. 7). The Oneida orthography adopted in this paper as well as our database is consistent with the one used in the Twatati Adult Oneida Language program.

2.3 Morphosyntactic Structure of Oneida Verbs

Figure 2 below by Michelson and Doxtator (2002) illustrates the structure of Oneida verbs. At the very core of each Oneida verb there is the verb root. Then, to form the verb base, the optional reflexive, or reciprocal prefix, as well as the incorporated noun can be attached before the verb root while the optional derivational suffix can be attached after it (Michelson and Doxtator, 2002).

Next, to form the verb stem, the aspect suffix is required to be attached after the verb base. Attachment of the pronominal prefix before the verb is also obligatory as this morpheme encodes information about the arguments in the utterance. Then, before the pronominal prefix, the prepronominal prefix can be added optionally to enrich the expression with information such as dualic, translocative, cislocative, repetitive, or negative (Michelson and Doxtator, 2002, p. 27). However, the presence of the prepronominal prefix can also be obligatory as they are required by a small number of verb roots or aspects due to their special properties (Michelson and Doxtator, 2002).

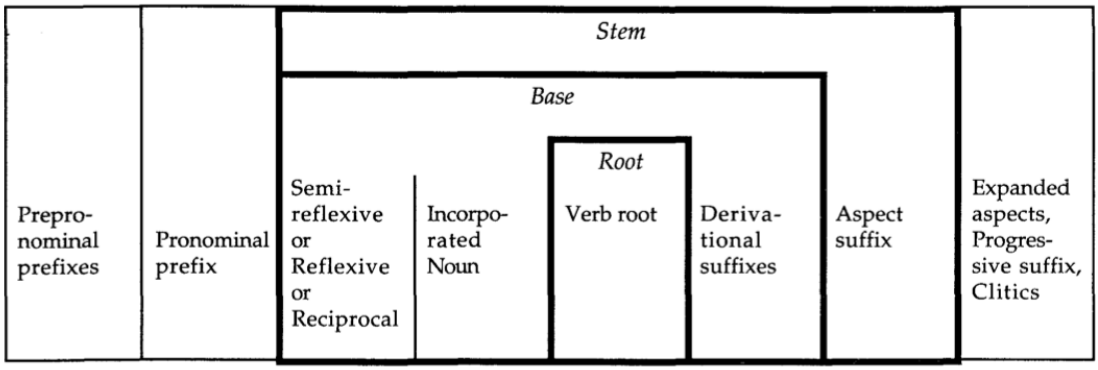

Figure 2: Positions of elements that form Oneida verbs (Michelson and Doxtator, 2002, p. 14)

[1] Below is the list of abbreviations used for the glossing of the examples: 1: first person; 2: second person; 3: third person; SG: singular; DU: dual; EXCL: exclusive; F: feminine; FI: feminine indefinite; FZ: feminine zoic; HAB: habitual; M: masculine; N: neuter; NEG: negation; PL: plural; PPR: pronominal prefix; PREPPR: pre-pronominal prefix; PTCL: particle; Q: question particle; STAT: stative

[2] The length of Oneida utterances can vary greatly and can be difficult to define. It can be as simple as one word or as complex as multiple sentences that can qualify as a paragraph in other languages such as English (Michelson et al., 2016, p. 9).

[3] Among the analysis of different authors, there exist minor disagreements regarding whether certain sounds are classified as phonemes or allophones in Oneida. The phonetic data presented in this paper are not directly taken from one specific source but are adaptations of analyses from four different works.

At the moment, the verb conjugator only includes the three morphemes required for forming an independent utterance: the verb root, the aspect suffix, and the pronominal prefix. In the following subsections, we will present each of the three morphemes in detail. Further information about each component of an Oneida verbal complex/utterance can be found in Lounsbury (1976), Michelson and Doxtator (2002), and Michelson et al. (2016).

2.3.1 Verb Roots

Oneida verbs can have various functions: describe events and states, express kinship, refer to entities, and describe the properties that are often expressed by adjectives in languages like English (Michelson et al., 2016, p. 343). These verbs can be categorized into two classes based on their meanings and properties: *active* verbs and *state* verbs (Michelson et al., 2016, p. 343). Active verbs can occur in all aspects: habitual, stative, and punctual, as well as in the imperative form; contrastively, state verbs can only occur in the stative aspect (Michelson et al., 2016, p. 343).

2.3.2 Aspect Suffixes

The three aspects of Oneida are distinct from the ones of English. Habitual aspect describes repeated or ongoing actions and can be equivalent to "do something", "keep doing something", or "doing something" in English (Michelson and Doxtator, 2002, p. 19). Stative aspect describes actions or events that have happened in the past but their effects last after their completion. This aspect is equivalent to "have done something", or in some cases "doing something", or "it is done" in English (Michelson and Doxtator, 2002, p. 19). The punctual aspect describes the events that happened as a single occurrence. This aspect suffix requires a modal prepronominal prefix to be attached at the front of the word, while these modal prepronominal prefix also change the interpretation of the event (Michelson and Doxtator, 2002, p. 19). Even though state verbs and active verbs can both occur in the stative aspect, the state verbs do not share active verb's freedom of appearing in any of the other aspects and they are also typically intransitive (Michelson and Doxtator, 2002, p. 27). However, the forms of the stative aspect suffixes are the same for both types of verbs.

A subset of active verbs are motion verbs, which depict the way movement occur (Michelson et al., 2016; Michelson and Doxtator, 2002). Furthermore, a fourth aspect, the intentive aspect, exists only for this type of verbs, for instance, **intentive**: *katawʌ́·ne'* 'I'm going to swim, I'm going to bathe', compared to **habitual**: *katawʌ́·nehse'* 'I go swimming' (Michelson and Doxtator, 2002, p. 22). Exceptions and variations to these rules exist and are determined by the property of the verb as well as how natural the meaning is interpreted by the speakers (Michelson and Doxtator, 2002, p. 19, 21). Michelson and Doxtator (2002) classifies all of the Oneida aspects into different classes and labels each of them with a unique combination of letters and numbers. This classification is followed in our database.

2.3.3 Pronominal Prefixes

Oneida pronominal prefixes encodes the person, number, gender, and inclusivity information about the agent and/or the patient. In Oneida, there are three persons: first, second, and third, and the first person differentiates exclusive from inclusive. There are three numbers: singular, dual, and plural. Oneida has four genders: masculine, feminine zoic (FZ), feminine indefinite (FI)[4], and Neuter. Oneida pronominal prefixes can be divided into three classes based on the semantic animacy of the arguments (Koenig & Michelson, 2015). The Oneida and Kanyen'kéha immersion programs colour code these three classes into three different colours to help students distinguish them more efficiently. The method was first invented by the late Kanatawakhon (David) Maracle. In this paper as well as the database, we follow the same practice: if there are two animate arguments in the utterance, a purple pronominal prefix is used; if only one animate argument occurs in the utterance, and the animate argument is the agent, a red pronominal prefix is used; if the animate argument is the patient, a blue pronominal prefix is used (Twatati, 2017). This system is illustrated by examples (1-3) below:

1) Thikʌ́ kʌ sʌ́ha' swanú·wehse'?
thikʌ kʌ sʌha' swa- -nuhwe'-se'
that Q more 2PL.A like-HAB
'Do you all like that (one thing) more?'

2) Yah úhka' tehuwanú·wehse'.
yah uhka' te- huwa- -nuhwe'-se'
nobody PREPPR 3SG/PLFI>3SGM like-HAB
'No one likes him.'

3) Yah kʌ tehonanúhte' oh náhte' kuwa·yáts?
yah kʌ te- hon- -anuhte-'
NEG Q PREPPR 3PLM.P know-HAB
oh nahte' kuwa- -yat-s
PTCL what PLFI>3SGFZ call-STAT
'Don't they(mf) all know what they call (it)?'

(adapted from Twatati, 2017)

The verb root *-nu·weh-* 'like' can take up to two animate arguments. In (1), because there is only one animate argument, which is the agent, a red pronominal prefix *swa-* is used. In (2), we see the same verb root *-nu·weh-* 'like', however, there are two animate arguments in this utterance, the 'liker' and the 'being liked'. In this case, a purple pronominal prefix

[4] According to Abbott (1984, p. 126), speakers usually make their decisions of using FI or FZ based the following 6 factors: (1) indefiniteness; (2) animacy; (3) humanness; (4) size and gracefulness; (5) age; and (6) special relationship (Abbott, 1984, p.126). There exist a small number of examples where a female person or otherworldly being is referred to with both FI and FZ genders (Michelson, 2015). Overall, referring to inanimate entities with FI is illegal while FZ can refer to animals, human beings, or inanimate objects that "acquire the animate property of locomotion" (Abbott 1984; Michelson, 2015, p. 291).

huwa- is used. In (3), the verb root -*anuhte*- 'know' requires the sentence to have only one animate argument, which is the 'knower', and the 'knower' must be the patient. In other words, the knowledge comes to the knower(s). Therefore, a blue pronominal prefix must be used in this case.

This pattern applies to most verbs in Oneida. However, in many cases, semantics alone cannot determine which pronominal prefix is selected (Koenig and Michelson, 2015, p. 5-7). For verbs with a single animate argument, the selection between blue or red pronominal prefixes can also be determined by the property (such as aspect) and meaning of the verbs (Koenig and Michelson, 2015, p. 8-9; Michelson et al., 2016).

In addition to these three classes, each Oneida pronominal prefix has five allomorphs, determined by the initial segment of the verb stem to which they are attached (Twatati, 2017). If the verb stem begins with a consonant, then a C stem allomorph is used. If the verb stem begins with one of the six Oneida vowels, then one of the four vowel stem allomorphs is used (A stem if the verb stem begins with /a/; I stem if the verb stem begins with /i/; E/ʌ stem if the verb stem begins with /e/ or /ɛ̃/; O/U stem if the verb stem begins with /o/ or /ũ/)[5]. Furthermore, if a prepronominal prefix is attached before the pronominal prefix, it may also cause phonological variations to the initial segment of the pronominal prefix. For instance, when the prepronominal prefix *te*- is attached, the initial segment *l* of the pronominal prefixes must become *h*, so *te*- plus *luwa*- and *lon*- surface as *tehuwa*- and *tehon*- (Twatati, 2017).

2.4 Oneida Phonology

The phonological system of Oneida intricately intertwines with its morphology[6]. For instance, the most fundamental stress pattern of Oneida is penultimate stress; however, exceptions created by processes such as shifting and lengthening arise based on the syllabic structure (Michelson, 1988). Meanwhile, changes to the syllabic structure of a word are directly influenced by the processes of affixation. These rules, together with other phonological intricacies such as vowel epenthesis and laryngeal lengthening, make stress placement complex, sometimes even leading to native speakers disagreeing on stress placement for specific words (Michelson, 1988). Due to these complications, Gramble is not yet capable of automatically attributing prosodic features to the conjugated words. Therefore, we are leaving stress placement to future work.

3. Motivations and Objectives

In the previous sections, we have briefly presented the word, sentence, and sound structures of Oneida, from which it is easy to see the prominent differences between learners' L1 English and L2 Oneida. Imagine when an Oneida learner is trying to conjugate a verb root with the correct pronominal prefix and aspect suffix, they must make multiple decisions from a number of options within a very short time. Below is a step-by-step illustration of the questions they have to ask and answer in order to conjugate the verb in the most basic situation where the verb roots are conjugated with only two affixes, the pronominal prefix and the aspect suffix:

1. What type of verb is it (Active/Motion/State)?
2. What aspect is the expression in (Habitual/ Stative/Punctual/Intentive)?
3. What's the phonological environment of the suffix? Any variations triggered?
4. How many animate arguments are there? /What class of pronominal prefix does the verb/aspect require?
5. What are the person, number, gender, and inclusivity features of the participant(s)?
6. What is the initial segment of the verb stem? /Which allomorph should be used?
7. What additional phonological variations are triggered once the affixes are attached?
8. Where should stress be assigned? Does the assignment of stress cause further phonological variations?
9. Is there any additional variation caused by factors such as morphological, lexical, cultural, or conventional requirements?

All of these complex processes are for selecting only two affixes, while utterances in everyday conversations often contain four or five morphemes. Every additional morpheme can cause the complexity of the utterance to multiply, which makes Oneida verb conjugation very challenging for beginner learners to acquire. This means having lots of materials for learners to practice forming verbs and to check their answers is crucial. However, recording each of these combinations in a textbook or a dictionary would be very not practical if not impossible. It is estimated that it would take approximately 20 years for someone to type out each of the possible conjugations of Kanyen'kéha (Anna Kazantseva, p.c.). With the significant overlap of core vocabulary and morphosyntactic structures between Kanyen'kéha and Oneida (Julian, 2010), we can expect that creating a textbook of Oneida verb conjugation will take someone at least two decades as well, and this is not to mention the likelihood that we as human beings would make frequent errors in such repetitive work.

Given the lack of opportunities of practicing conjugating verbs outside of the classroom with fluent speakers and the lack of reliable or complete textbooks to look up the answers, Oneida students report that this is one of the most challenging parts of their process of acquiring the language. These obstacles are not unique to Oneida students, the students of the Onkwawenna Kentyohkwa adult

[5] Detailed explanations and examples can be found in Lounsbury (1976), Michelson and Doxtator (2002), and Michelson et al. (2016).

[6] Detailed phonological analyses of Oneida can be found in Michelson (1983), Michelson (1988), and Lounsbury (1942).

Kanyen'kéha immersion school have echoed this feedback. This feedback has led to the development of the digital Kanyen'kéha verb conjugator, Kawennón:nis[7] (Kazantseva et al., 2018), as a collaboration between Onkwawenna Kentyohkwa and National Research Council of Canada. On the backend, Kawennón:nis uses a handwritten finite-state transducer (FST), and on the front-end uses WordWeaver[8], an Angular-based JavaScript interface for verb conjugation. It is online and freely available to the general public and fully owned by Onkwawenna Kentyohkwa.

With Kawennón:nis, student can choose the verb roots, the agent and/or patient, and the tense/aspect for the verb conjugator to produce the corresponding conjugated form. Students can also adjust the settings and Kawennón:nis will highlight the different morphemes with different colours or visually illustrate the steps of how a verb is conjugated. These functions empower students by allowing them to easily and accurately look up conjugations which they might be unfamiliar or uncertain. On the other hand, some students find practicing drills beneficial for their memorization of the conjugated forms. These students can also use Kawennón:nis to create and download practice materials tailored to their interests and needs. Students from the Onkwawenna Kentyohkwa immersion school along with independent learners have attested that Kawennón:nis greatly facilitate their learning of morphosyntactic language structures of the language.

The curriculum of the Twatati Adult Oneida Language program is developed based on the curriculum of the Onkwawenna Kentyohkwa adult Kanyen'kéha immersion school. Both curricula follow the rootword method created by the late Kanatawakhon (David) Maracle. Therefore, we believe an Oneida version of the verb conjugator will serve as a greatly beneficial tool for Oneida learners as well. We presented the idea as well as Kawennón:nis to key members of the Twatati committee: Nancy George, Ursula Doxtator, and Tania Granadillo. The proposal was met with enthusiasm from the committee members, and they also expressed commitment to support our endeavor as collaborators.

Drawing upon the timelines from previous verb conjugator projects, it appears that these projects often span multiple years before reaching a point of "completion". Given the nature of a verb conjugator, its development will remain an ongoing process. The database can be continually optimized and expanded, with new verb roots regularly incorporated. So far, we have only completed a pilot version of the conjugator, and we will continue to expand and refine this project in the coming years. Once all collaborators and participants are in agreement that the conjugator is in a satisfactory condition, the ownership together with the responsibility of maintaining and enhancing the verb conjugator will be passed to Oneida teachers or enthusiasts from the Oneida Nation of the Thames. While we plan to step back from the front lines of development at this stage, we will remain committed to providing support and assistance when required to ensure the success and longevity of this valuable linguistic tool.

3.1 What is Gramble?

During the development of Kawennón:nis and similar tools, that team encountered difficulty in using the XFST and LEXC languages (Beesley and Karttunen, 2003) to develop grammars for complex grammars in a multi-skilled team. In particular, collaborating subject matter experts often had difficulty understanding the code, which limited their direct participation beyond early stages and caused concern about the eventual product hand-over.

During development, the team began developing an in-house toolkit as an XFST replacement, named Gramble (Littell et al., 2024), which combines a richer formalism (based on *n*-tape automata rather than 2-tape automata) with a simpler syntax. Unconventionally, Gramble uses a tabular syntax rather than plaintext, and can be programmed using a spreadsheet editor. By design, the tabular syntax is very similar to an ordinary spreadsheet, so that a typical knowledge worker can read and write the basics in a familiar environment, even without a computer science background.

We adopted their framework for similar reasons, both because the primary developer's background is in traditional linguistics rather than NLP, and also to better involve community experts in its development and continued maintenance.

3.2 Why Not Machine Learning?

There are several reasons we chose a rule-based system instead of a system based on a neural language model.

First, as mentioned earlier, Oneida does not currently have a large number of speakers, and the amount of digitized material is limited. As noted in Section 5, it is difficult to even find enough relevant data for evaluation, let alone training. This issue is faced by most Indigenous languages, which causes the majority of current technologies for Indigenous languages to be rule-based as well (Arppe et al., 2016; Littell et al., 2018; Kuhn et al., 2020).

Second, especially in very low data scenarios like this, there is a high risk of fabrication: the generation of seemingly-realistic outputs unrelated to the input. This is inappropriate in an educational reference tool. Most users will be learners, who have not yet mastered the language enough to detect spurious forms and could acquire these in place of genuine Oneida.

The final reason concerns the ability to fix incorrect outputs. A practical advantage of traditional, non-neural natural language generation (NLG) is that it is more straightforwardly fixed when the client encounters inappropriate outputs (Reiter, 2021). If the client found incorrect outputs in a neural NLG

[7] https://kawennonnis.ca/wordmaker

[8] https://github.com/nrc-cnrc/wordweaver

system, how exactly we would fix it is unclear. (Perhaps feeding it the corrected data during training would fix the problem, perhaps it would not; the bigger issue is that fixing the system becomes a research project of its own rather than a bugfix.)

In the big picture, a system based on human-written rules, that the community can access, read, and change, helps establish trust that the Oneida instructors are the ultimate decision makers regarding their language. Even if a neural model *did* have adequate accuracy here, adopting an unexplainable black-box model does not move us towards this greater goal.

4. Implementation

4.1 Resources Consulted for The Database

The data used to develop the grammar of the Oneida verb conjugator come from the following publicly available resources:

1. The curriculum of the Twatati Adult Oneida Language program.
2. The *Oneida English/English Oneida Dictionary* authored by Karin Michelson and Mercy Doxtator, based on the Oneida of the Oneida Nation of the Thames.
3. The book *A Comparative Study of Lake-Iroquoian Accent,* also authored by Karin Michelson.
4. The book *Glimpses of Oneida Life* authored by Karin Michelson, Norma Kennedy, and Mercy Doxtator.
5. The book *Oneida Verb Morphology* authored Floyd Lounsbury, based on analyses of Oneida stories told for the Works Progress Administration (WPA) project during the end of the 1930s and the beginning of the 1940s.

4.2 Describing the Grammar in Gramble

Due to the large size and complexity of the morphemes and allomorphs, the verb roots and affixes are listed in separate sheets and tables then joined together obeying the specified sequences and rules. First, for the table of verb roots, each verb is accompanied with specifications of their underlying form, their English translation, colour class, root class, aspect class, as well as any additional comments or notes about the form. In the database, stress is not marked on any of the morphemes unless it is required for creating the environment for certain phonological rules to take place. As explained earlier, affixation processes trigger relocation of the stress of the word in complex ways, and we have not yet managed to express this completely through phonological rules. Stress in Iroquoian languages is particularly difficult to capture computationally using rules alone (Anna Kazantseva and Akwiratékha' Martin, p.c.).

Next, the verb roots are attached to the aspect suffixes to form verb stems where the pronominal prefixes are attached later. Following the analysis and categorization by Michelson and Doxtator (2002), these aspect suffixes are divided into coded classes. Some of these morphemes exhibit allomorphic variations influenced by varying phonological environments. These variations are included in the table, separated by the pipe symbol " | ", however, we have not yet been able to make the database automatically select and attach the correct variation. As a result, verb roots that require these specific classes of suffixes have been temporarily omitted from the database. Next, attachment of the aspect suffix to the verb root creates changes to the phonological structure of the word which triggers variations. Therefore, specific phonological rules must be applied to the conjugated verb roots at this stage in order to create the correct environment for further phonological processes.

Then, on three separate sheets, all forms of the red, blue, and purple pronominal prefixes are listed respectively with the person, number, gender, and inclusivity features, as well as the classifications specified for each form. While the sheets for the red and the blue pronominal prefixes are relatively straightforward with around eighty lines long each, the combinations of different agent and patient cause the purple sheet to be over a thousand lines long. To keep the tables compact and easy to read, some of the portmanteau morphemes that encode multiple grammatical features are collapsed in the main table and broken down elsewhere in the sheet. Additionally, on each of these sheets, another set of tables are used to classify the verb stems, composed of verb roots and aspect suffixes, into the five stem types (C-stem, A stem, I stem, E/ʌ, and O/U stem). These tables ensure the accurate attachment of each verb stem with their respective allomorph of the pronominal prefixes. Once the pronominal prefix is attached, the phonological structure of each word changes again and additional phonological rules are specified in another table. Figure 3 below is a screenshot of part of the purple pronominal prefix sheet. Note that what it shows is the code itself instead of a representation of codes that were originally written in plain text.

A pilot version of the Oneida verb conjugator interface[9] (as shown by Figure 1) has been created for demonstration purposes during the meeting among collaborators of the project in November 2022.

purple_PPR=	table:	subject_base/subject	English/subject_translation	text	embed	object_base/object	English/object_translation
		1SG	I>	ku	C	2SG	>you
		1SG	I>	kni	C	2DU	>you_two
		1SG	I>	kwa	C	2PL	>you_all
		1SG	I>	li	C	3SGM	>he

Figure 3: Snippet of Gramble code describing purple (transitive) pronominal prefixes

[9] Our sincere gratitude to Delaney Lothian (Application Development Specialist of the ILT team) for helping us set up the user interface of the Oneida verb conjugator.

It remains restricted by a password until we receive a green light from our collaborator at the Oneida Nation of the Thames, signaling its readiness for public release.

5. Quantitative Evaluation

5.1 Methodology

We performed a basic evaluation of the verb conjugator by testing to see if the forms generated by the verb conjugator match with forms extracted from documented examples of conjugated verbs. Considering this is only the preliminary stage of the project, and the combination of multiple morphemes can easily result in a vast number of output data, we limited the forms included in the test to the most fundamental and simple cases: purple (transitive with animate agent and patient) pronominal prefixes, active-class roots, and the habitual aspect. 56 verb roots of all five stem types are included in the database (74 forms if duplications that reflect phonological variations are also counted) and 8 replacement rules (4 before the attachment of the pronominal prefix and 4 after) are specified based on the phonological rules of the language explained in Twatati (2017) and Michelson and Doxtator (2002). The combination of these purple pronominal prefixes with active and transitive verb roots as well as habitual aspect suffixes lead to the generation of 8475 unique forms of conjugated verbs.

The evaluation set contains 100 forms of conjugated examples extracted from the *Oneida-English/English Oneida Dictionary*. This is admittedly a small test set, so one should not put too much stock in the results, but this is an under-resourced language. For any given subset of the verb paradigm, there are simply a limited number of attested forms to which we can compare system outputs and finding them is labour-intensive. Since so few examples are attested, this evaluation focuses only on recall: ideally, every attested form should be generated, but even in the best-case scenario only a tiny fraction of generated forms will be attested.

Although this is not a machine-learning project, we adopt the evaluation paradigm in which evaluation data are randomly divided into "dev" and "test" sets, to simulate the performance of the system on unseen/future data. The results of "dev" are revealed to the author for error analysis and further development, whereas "test" was held out. (Eventually, of course, the team will look at this data and make sure they are handled properly; we do not want to put a system in front of students that we know is making errors. But for the purposes of *this* evaluation, temporarily holding back data ensures the programmer is writing rules that generalize, rather than achieving high accuracy by writing rules that fix specific incorrect forms in the test set.)

As mentioned in Section 4.2, the model currently generates prosodic information for only a subset of forms, so this preliminary evaluation disregards prosodic features and only evaluates accuracy on segmental material.

5.2 Results and Error Analysis

The system achieves a recall of 93% for both the "dev" data and the "test" data. A total of 44 forms are included in the "dev" group and 41 of them are correct. The three forms missing from the output of the Oneida verb conjugator are *lakenhlálhos* 'He keeps giving me his germs', **shako'tanʌwʌhslályo* 'He keeps whipping her', and *kheyahta'nawʌ́sta'* 'I dress her up warmly'.

Each of the three incorrect forms are caused by a different reason. In the case of *lakenhlálhos*, the phonological rule of Oneida requires an *e* to be inserted between the pronominal prefix and the verb root. However, description of the trigger of the insertion varies slightly between different resources. According to Michelson et al. (2016), "Prefixes that end in a consonant have variants with **e** after the consonant before stems that begin in **kh**, **sh**, **sk**, **sl**, **st**, **th**, **tsh**, **tsy,** or **ʔ**" (p. 348). Meanwhile, according to the curriculum of the Twatati program "The ***e*** (underlined and in italics) is used when root begins with **'** or double consonant. Example: sknú·wehse' but ske'nikú·lale'" (Twatati, 2017, p. 8). The rules we have added to the database include all of the cases described in Michelson et al. (2016), however they do not form the complete list of all of the possible triggers of the insertion. As we can see with the form *lakenhlálhos* 'He keeps giving me his germs', the *nh* consonant cluster triggers the insertions of *e* but is not mentioned in Michelson et al. (2016). Through further consultation of Michelson (1983), we discovered a rule that specify *e* is inserted before extra-syllabic consonants at stem-initial position or glottal stops (i.e., *ʌ-yu-atat-nha'-n'* becomes *ʌyutat**é**nhane'* 'she will hire her' after e-epenthesis before the verb stem -*nha'*-) (p. 225). Given additional time, we will consult a diverse range of sources in subsequent stages of the development of the verb conjugator to ensure the overall robustness and accuracy of the database.

Next, in the case of **shako'tanʌwʌhslályo*, this error is simply resulted from a typo we made when compiling the test set. Despite having double-checked the forms before the test, this form escaped our notice that an *s* is missing at the end of the word. The correct form should be *shako'tanʌwʌhslályos* instead, which in fact can be found among the output of the verb conjugator. This also demonstrates the advantage of having a digital verb conjugator and the problem mentioned earlier that humans are prone to making mistakes as the dataset gets bigger and bigger.

In the case of *kheyahta'nawʌ́sta'*, this form is composed of a pronominal prefix *khe-*, the verb root -*yahta'nawʌst-*, and the suffix -*ha'*. However, we can see that the initial *h* is missing in the conjugated form. Unlike other entries, this variation is not explained in the dictionary. After consulting Michelson (1983, 1988), we realized that we overlooked the rule that *h* is deleted when it occurs after a CC cluster. Once again, we will conduct additional in-depth consultations on Oneida grammar to refine and enhance the database.

For the “test” group, out of the 56 forms included, 52 of them are correct. As these incorrect forms are not visible to us at the moment, we are not certain what has triggered the errors. They could be caused by the similar issues as above, or they could be individual cases of variations that do not comply with the rules. Once the initial stage of the project is completed, we will closely examine these cases and discuss them with our consultants.

6. Qualitative Evaluation

Several unforeseen circumstances, including the far-reaching impact of the COVID-19 pandemic and the regrettable loss of elder speakers, have hindered the consistent scheduling of regular meetings with participants as initially intended. So far, we have had three meetings: two held online in October and November 2022, and one in-person at the classroom of the Twatati program in July 2023. During the virtual meetings, we engaged in discussions with two L2 learners and coordinators of Oneida language programs who offered us invaluable feedback based on their personal experiences and observations. The third gathering marked a significant milestone: it was the first in-person gathering of the Twatati committee since the COVID-19 pandemic. This meeting was attended by members of the Twatati committee and participants of their Master Apprentice program. Among them there are six L2 learners and one native speaker of Oneida. Here, we provide a comprehensive overview of feedback, inquiries, and concerns raised during these meetings, these responses will help shape the trajectory of this project.

(i) The participants suggested that the verb conjugator should also include audio recordings of each conjugated form to demonstrate their accurate pronunciation. It is often the case that a student has made considerable progress in learning how to accurately conjugate the verbs, however, their pronunciation of the verbs still needs improvement. The L1 speaker shared her observation that the speech of L2 learners often sound unnatural and “choppy”. This issue could potentially cause the learners’ speech difficult to comprehend for native speakers. Therefore, having audio representations will be especially beneficial for helping students grasps the suprasegmental features which are hard to interpret based on their written representation alone but are crucial for achieving advance proficiency.

To fulfill this request, two approaches can be taken: 1) making recordings of speakers’ pronunciation of each form and attach them to the written form correspondingly; and 2) using technologies such as speech synthesis to automatically generate audio representations of each form. Each method has their benefits and costs in terms of efficiency and authenticity, decisions of which approach to adopt will be made in consultation with participants and each collaborator of the project in the future stages.

(ii) One of the main topics of the discussions is about the ownership of the verb conjugator. We have assured members of the Twatati committee as well as the participants from the Oneida community that we will not claim ownership to any of the language data. Once the project is completed, the full ownership will go to the Oneida Nation of the Thames. However, the question of which individual or organization of the community should be the optimal owner of the verb conjugator remains to be decided.

(iii) The participants also asked about the ongoing expenses associated with the website's upkeep and its hosting platform. Since Gramble and the user interface operate entirely on the client side, this means that it is not necessary to provision a back-end server, except to serve a static webpage. Currently, the user interface is hosted on a free GitHub page. If the future owner of the conjugator chooses to migrate the website to another hosting solution or transition it into a mobile app, such changes can be achieved with manageable effort and expense. We anticipate that such changes will not pose substantial challenges in the long run.

(iv) The participants expressed enthusiasm about integrating more technology into the teaching and learning process of Oneida. They believe that the younger generation, who are excited about the latest technology, would be greatly motivated to engage with the learning materials and to use the language more often if more digital tools and resources are introduced. The participants trust that the verb conjugator will make significant contributions to the revitalization of the Oneida language.

7. Discussion

In this section, we discuss the limitations of the project as well as the plan for tackling these issues in the future.

7.1 The Issue of Overgeneralization

Digital tools such as verb conjugators often face the challenge of overgeneralization, that the complete list of possible combinations generated by the verb conjugator exceeds the number of forms actually used within the community by native or fluent speakers. The software can only follow the rules from the database strictly and generate all the possible combinations. However, we know that real-world languages are much more complex than just following grammar rules. Different expressions can have additional meanings in specific communities or meanings that make them inappropriate for many contexts. Member of Oneida and Kanyen’kéha language programs as well as researchers of the languages often share stories of how learners conjugate forms that are technically correct based on the grammar, but have a ‘weird’, ‘funny’, or sometimes ‘offensive’ meaning within the community. This weakness of the verb conjugator can potentially cause issues within the language programs. Users might learn the forms generated by the verb conjugator while being unaware that they are ungrammatical or unnatural. Not only can these inaccurate or inappropriate forms negatively influence the learning outcome of individual learners but also

their peers as well. With the lack of native speakers in the community these Oneida learners may not be corrected in time before they start using and spreading these newly acquired inaccurate forms among each other.

To tackle this issue, we will compare all of the forms generated by the verb conjugator with existing forms found in documentation materials. Any form that has not been previously recorded will be highlighted and accompanied with a warning message indicating that these forms need to be used with caution. At the meantime, we will consult with native or fluent speakers to validate the accuracy of the forms that have not been previously recorded. This step is essential as input from speakers will enrich the quality, quantity, and authenticity of the content of the database. Users will also be encouraged to report any form that they discover as inaccurate or inappropriate. Any such form will be eliminated or flagged manually. Once the responsibility of maintaining and improving the verb conjugator is passed on to the Oneida community, access to the database can be shared among community members so that all users will be involved in the task of solving this issue.

7.2 Lack of Complexity

Most verb conjugators are limited by the complexity of the morphological structures of polysynthetic languages. As mentioned in section 2.3, the Oneida conjugator currently contains only the verb root, the pronominal prefix, and the aspect suffix. Complex yet common and essential elements such as noun incorporation are not included due to the fact that they multiply the output forms and trigger numerous additional (morpho)phonological variations. Avoiding complications through omitting advanced features can significantly limits the effectiveness and benefits of the tool. This creates a dilemma that the inclusion of additional complex structures leads to sacrifice of accuracy or speed while pursuit of accuracy may require sacrifice of complexity.

To tackle this issue, it is essential to seek advice from users of the previous verb conjugators and the current Oneida conjugator to learn if they find the inclusion of more complex structures useful. Due to the lack of studies on adult L2 acquisition of polysynthetic languages, and languages of the Iroquoian family more specifically, we are yet uncertain about the processes and stages of morphological acquisition of adult Oneida learners. If students have already mastered the rules of Oneida verb conjugation by the time they are utilizing complex phrases, then the majority of the users of the tool will be beginner learners. Excluding the more complex structures might be more beneficial as it eliminates unnecessary confusions. However, if the process of acquiring the conjugation rules is still on-going as the learner move on to using advanced structures, then inclusion of the complex structures is necessary for it to remain useful for advanced learners.

7.3 Suprasegmental Features

To date, the integration of suprasegmental features into the verb conjugator remains an unresolved challenge. The intricate variations stemming from Oneida's phonological rules prevent Gramble from automatically attributing prosodic features with a satisfactory accuracy rate. The new Kanyen'kéha verb conjugator for the Kahnawà:ke (Eastern) dialect being developed with Gramble also encountered this challenge. While they have made significant progress, no solution has yet been discovered that addresses this problem. However, Kawennón:nis, which is developed through WordWeaver have achieved remarkable success in tackling this issue. The vast majority of the forms generated by Kawennón:nis reflect accurate markings of prosodic features. This will be one of the primary emphases of the forthcoming steps of our project.

8. Conclusion

In this paper, we presented the work we have completed in collaboration with the Twatati Adult Oneida Language program team for the preliminary stages of developing an Oneida verb conjugator. The evaluation of the output of the verb conjugator shows promising results. Although there are several issues that remain to be tackled, the benefit of the Oneida verb conjugator still significantly outweighs its drawbacks. As we move forward, the journey of refinement and advancement of various aspects of the project is ongoing to ensure its continued alignment with the language revitalization goals of the Oneida community. The active participation of Oneida learners and native speakers will remain indispensable. Their insights, feedback, and continual engagement are key for ensuring the conjugator to be accurate and effective in fostering linguistic growth within the Oneida community.

9. Ethics Statement

Despite numerous previous studies in the field of second language acquisition, the majority of them concentrate on widely spoken languages like English and Spanish (Miyashita and Chatsis, 2013). The inherent differences between the natures and systems of these languages and the polysynthetic Indigenous languages would cause these research findings and pedagogical advice to be not applicable to Oneida. Furthermore, second/foreign language programs and language revitalization programs diverge in multiple aspects, such as the goals, needs, expected outcomes, and learners' motives (Grenoble, 2009; Hinton, 2011). Concurrently, development of physical or digital language learning tools as well as language processing technologies have been booming in recent decades. Regrettably, these materials predominantly cater to languages of European origins or languages with large corpora, and few of them address polysynthetic Indigenous languages or meet the needs of Indigenous communities (Arppe et al., 2016). Moreover, the concept of digital verb conjugators is not novel, but they are often underrepresented in academic

literature, inadequately documented, or are not easily accessible by the public.

This project is a good example of the intersection of technology, linguistics, and Indigenous knowledge. It follows the community-based language research model proposed by Czaykowska-Higgins (2009) and the model of true collaborative fieldwork proposed by Leonard and Haynes (2010). It also provides a model that can be adapted and expanded for research projects on other Indigenous languages for and with other Indigenous communities facing similar challenges. This project contributes to a cumulative knowledge base in the field and supports continuous improvement in the development of technologies for Indigenous languages.

This Oneida verb conjugator itself will become a valuable tool that will contribute to the creation of more fluent Oneida speakers with shorter time and better efficiency and consequently support the restoration of the intergenerational transmission of the language. In addition to participants of the Twatati Adult Language program, this tool will also be of help to Oneida learners associated with other programs or studying on their own.

Meanwhile, some members of the Twatati committee have expressed their worries during our meetings. During the past few centuries, there has been a history of exploitation of traditional knowledge of Indigenous communities by unethical research projects, business cooperation, and government organizations; moreover, there are also the "parachute researchers" who enter Indigenous communities to collect data for their research but fail to return the generosity and assistance they've received and do not benefit the communities in return (Bradley and Bradley, 2019; Czaykowska-Higgins, 2009). In acknowledging the power, strength, resilience, and autonomy of the Oneida people, it is crucial to avoid depicting them solely as victims. However, the troubling historical events have led to a lack of trust towards outsiders among many community members and might cause them to unwelcome the verb conjugator. This concern once again reminds us that it is vital to adopt an approach that respects the history, culture, and autonomy of the Indigenous communities. At any point of this research, we must ensure all collaborations are built on a foundation of reciprocity.

10. Acknowledgement

With deep appreciation, we extend our gratitude to the members of the Twatati Committee: Nancy George, Ursula Doxtator, and Tania Granadillo, as well as the other members of the Gramble team at NRC: Akwiratékha' Martin, Anna Kazantseva, Darlene Stewart, Delaney Lothian, and Roland Kuhn for their support and input. We are grateful for Karin Michelson for her suggestions and feedback. We are also indebted to all those who have contributed to this project or helped us during this journey. Without them this project would not have been possible. Furthermore, we would like to thank the Linguistics Department of the University of Toronto for funding this project.

11. Bibliographical References

Abbott, C. (1984). Two feminine genders in Oneida. *Anthropological Linguistics, 26*(2), 125–137.

Abbott, C. (2016). Contact and change in Oneida. In A. L. Berez-Kroeker, D. M. Hintz, & C. Jany (Eds.), *Language contact and change in the Americas: Studies in honor of Marianne Mithun* (Vol. 173, pp. 167-188). John Benjamins Publishing Company.

Abbott, C., & Metoxen, L. (2012). Oneida language preservation. *Wisconsin Magazine of History, 96*(1), 2–15.

Antimirov, V. (1996). Partial derivatives of regular expressions and finite automaton constructions. *Theoretical Computer Science, 155*(2), 291–319.

Arppe, A., Lachler, J., Trosterud, T., Antonsen, L., & Moshagen, S. N. (2016). Basic language resource kits for endangered languages: A case study of Plains Cree. In C. Soria, L. Pretorius, T. Declerck, J. Mariani, K. Scannell, & E. Wandl-Vogt (Eds.), *CCURL 2016: Collaboration and Computing for Under-Resourced Languages: Towards an Alliance for Digital Language Diversity (LREC 2016 Workshop)* (pp. 1-8). European Language Resource Association.

Beesley, K. R., & Karttunen, L. (2003). *Finite state morphology*. CSLI Publications.

Bradley, D., & Bradley, M. (2019). *Language endangerment* (1st ed.). Cambridge University Press.

Brzozowski, J. (1964). Derivatives of Regular Expressions. *Journal of the ACM, 11*(4), 481–494.

Czaykowska-Higgins, E. (2009). Research models, community engagement, and linguistic fieldwork: Reflections on working within Canadian Indigenous communities. *Language Documentation & Conservation, 3*(1), 15-50.

Eberhard, D. M., Simons, G. F., & Fennig, C. D. (Eds.). (2023). Oneida. In *Ethnologue: Languages of the World* (Twenty-sixth edition). SIL International.

Grenoble, L. (2009). Linguistic cages and the limits of linguists. In J. Reyhner & L. Lockard (Eds.), *Indigenous Language Revitalization: Encouragement, Guidance & Lessons Learned* (pp. 61-69). Flagstaff: Northern Arizona University.

Hinton, L. (2011). Language revitalization and language pedagogy: new teaching and learning strategies. *Language and Education*, *25*(4), 307–318.

Julian, C. (2010). *A history of the Iroquoian languages* [Doctoral dissertation, University of Manitoba]. ProQuest Dissertations Publishing.

Kazantseva, A., Owennatekha, Ronkwe'tiyóhstha, & Pine, A. (2018). Kawennón:nis: the Wordmaker for Kanyen'kéha. In *Proceedings of the Workshop on Computational Modeling of Polysynthetic Languages* (pp. 53-64). Santa Fe, New Mexico, USA: Association for Computational Linguistics.

Koenig, J.-P., & Michelson, K. (2015). Invariance in argument realization: The case of Iroquoian. *Language, 91*(1), 1–47.

Kuhn, R., Davis, F., Désilets, A., Joanis, E., Kazantseva, A., Knowles, R., ... & Souter, H. (2020). The Indigenous Languages Technology project at NRC Canada: An empowerment-oriented approach to developing language software. *Proceedings of the 28th International Conference on Computational Linguistics* (pp. 5866–5878). International Committee on Computational Linguistics.

Leonard, W. Y., & Haynes, E. (2010). Making "collaboration" collaborative: An examination of perspectives that frame linguistic field research. *Language Documentation and Conservation, 4*, 268-293.

Littell, P., Kazantseva, A., Kuhn, R., Pine, A., Arppe, A., Cox, C., & Junker, M.-O. (2018). Indigenous language technologies in Canada: Assessment, challenges, and successes. *Proceedings of the 27th International Conference on Computational Linguistics* (pp. 2620–2632). Association for Computational Linguistics.

Littell, P., Stewart, D., Davis, F., Pine, A., & Kuhn, R. (2024). Gramble: A tabular programming language for collaborative linguistic modeling. *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024).* ELRA Language Resources Association and the International Committee on Computational Linguistics.

Lounsbury, F. G. (1942). *Phonology of the Oneida language* [Master's thesis, University of Wisconsin-Madison]. ProQuest Dissertations Publishing.

Lounsbury, F. G. (1976). *Oneida Verb Morphology*. Human Relations Area Files Press.

Michelson, K. (1983). A comparative study of accent in the Five Nations Iroquoian languages (Mohawk, Oneida, Onondaga, Cayuga, Seneca) [Doctoral dissertation, Harvard University]. ProQuest Dissertations Publishing.

Michelson, K. (1988). *A comparative study of Lake-Iroquoian accent*. Kluwer Academic.

Michelson, K. (2015). Gender in Oneida. In M. Hellinger & H. Motschenbacher (Eds.), *Gender Across Languages, Volume 4* (pp. 277-301). John Benjamins Press.

Michelson, K., Doxtator, M. A., & Kennedy, N. (2016). *Glimpses of Oneida Life*. University of Toronto Press.

Miyashita, M., & Chatsis, A. (2013). Collaborative development of Blackfoot language courses. *Language Documentation & Conservation, 7*, 302-330.

O'Grady, W. D., Archibald, J., & O'Grady, W. D. (2016). *Contemporary linguistic analysis: an introduction* (W. D. O'Grady & J. Archibald, Eds.; Eighth edition.). Pearson Canada.

Reiter, E. (2021) Challenging NLG datasets and tasks. Mar. 4, 2021. https://ehudreiter.com/2021/03/04/challenging-nlg-datasets-and-tasks/

Statistics Canada. (2022). Census Profile, 2021 Census of Population. Statistics Canada Catalogue no. 98-316-X2021001. Ottawa. Released October 26, 2022. https://www12.statcan.gc.ca/census-recensement/2021/dp-pd/prof/index.cfm?Lang=E

UNESCO. (2009). *UNESCO Atlas of the World's Languages in Danger*.

12. Language Resource References

Michelson, K., & Doxtator, M. (2002). *Oneida-English/English-Oneida dictionary*. University of Toronto Press.

Twatati Adult Oneida Language Program. (2017). *Ukwawʌ·ná· Kʌtyo·kwa' 1st Year Program*.