Empowering Oneida Language Revitalization: Development of An Oneida Verb Conjugator

Yanfei Lu, Patrick Littell, Keren Rice

LREC-COLING 2024

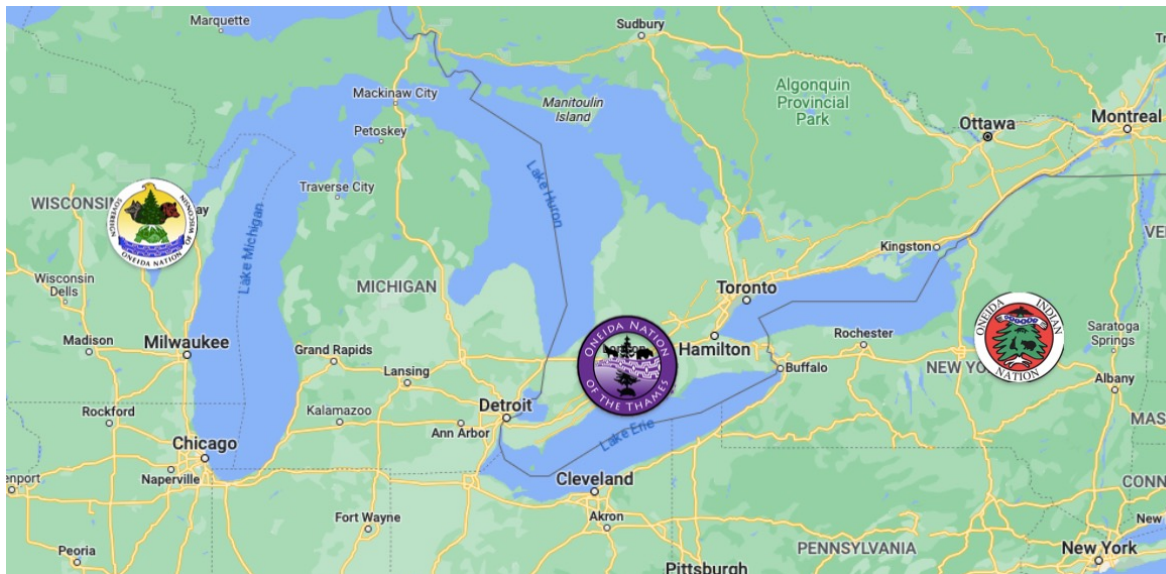# Background

Language of the Oneida People

Image 1. Locations of the Three Oneida Communities (Created based on Google Map)

3 main locations:

- Oneida Nation of the Thames near London, Ontario

- Oneida reservation near Green Bay, Wisconsin

- Oneida reservation near Syracuse, New York

(Michelson & Doxtator, 2002)

- 45 native speakers in Canada (Statistics Canada 2022)

- The *UNESCO Atlas of the World's Languages in Danger* marks Oneida as **critically endangered** (UNESCO 2009)

Iroquoian Language Family

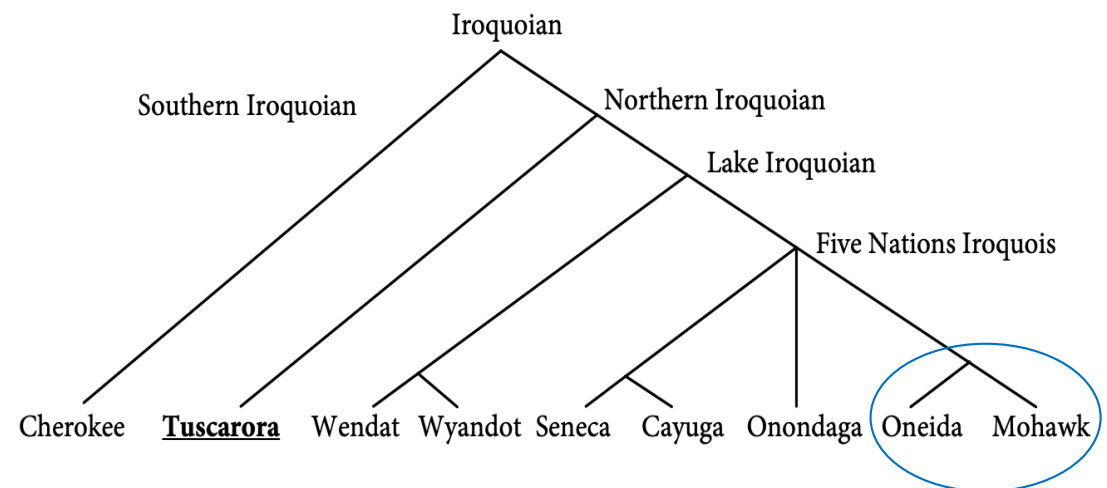- Related most closely to Kanyen'kéha (Mohawk)

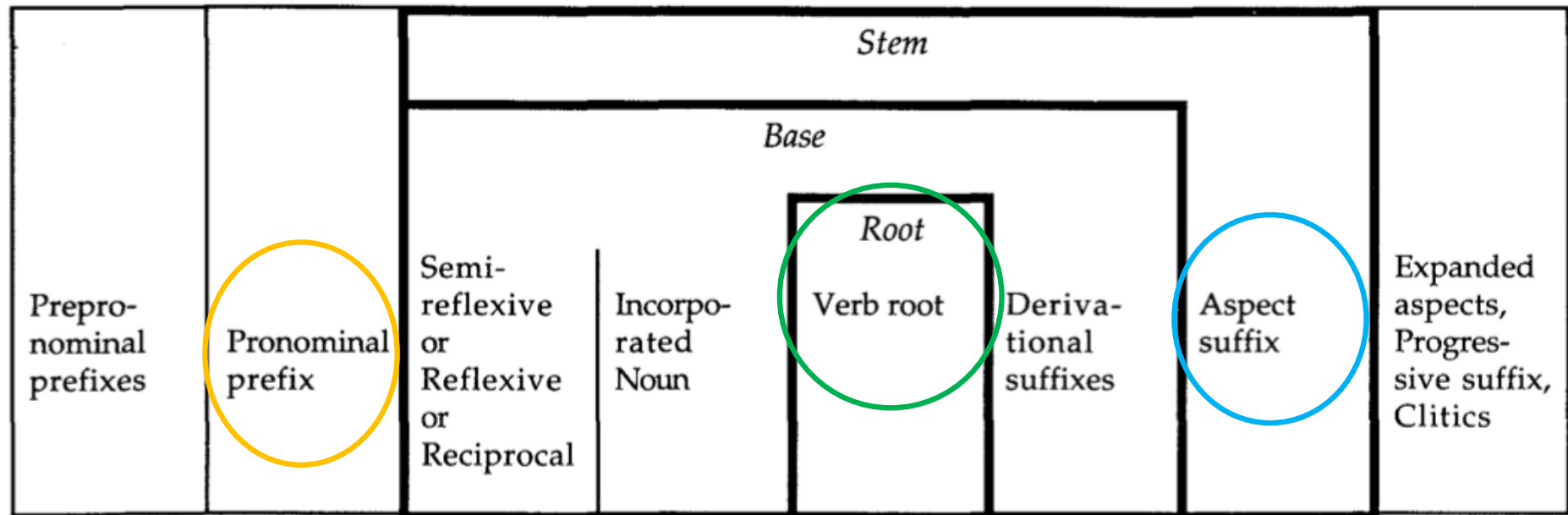Figure 1. Iroquoian Languages of Modern Day (Mithun 2013)

Polysynthetic

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

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

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

(adapted from Twatati, 2017)

Structure of Oneida Verbs

Michelson and Doxtator (2002, p. 14)

- Questions learners have to ask and answer to conjugate Oneida verb roots 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 (transitive/agent/patient)?

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 of pronominal prefix 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?
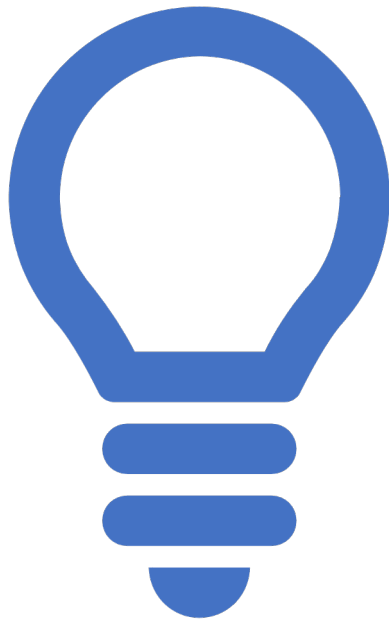
Motivation and methods
of the project

Motivation of the project

- Profound structural distinctions can lead to considerable challenges for adult learners
- The small population of native and fluent L2 speakers means insufficient opportunities for practicing Oneida
- Students report that learning how to conjugate verbs is one of the biggest challenges

Building the Digital Conjugator

What's the goal?	How to get there?
• Contributing to the revitalization and reclamation of the Oneida language through developing language learning tools dedicated for adult L2 Oneida learners	• Collaboration with the Twatati Adult Oneida Language program • Incorporating tools developed by the team of the Indigenous Languages Technology (ILT) project of NRC • Following a community-based research model (Czaykowska-Higgins, 2009)

Kawennón:nis

- Kanyen'kéha verb conjugator
  - Onkwawenna Kentyohkwa adult Kanyen'kéha immersion school
  - Indigenous Languages Technology (ILT) project of the National Research Council Canada (NRC)
- Owned by Onkwawenna Kentyohkwa

https://kawennonnis.ca/wordmaker

- Twatati Adult Oneida Language program
  - Oneida nation of the Thames
  - The curriculum is developed based on the curriculum of the Onkwawenna Kentyohkwa adult Kanyen'kéha immersion school

Community-Based Research

Community-Based Research

"Research that is **on** a language, and that is conducted **for, with**, and **by** the language-speaking community within which the research takes place and which it affects. This kind of research involves a collaborative relationship, a partnership, between researchers and (members of) the community within which the research takes place."
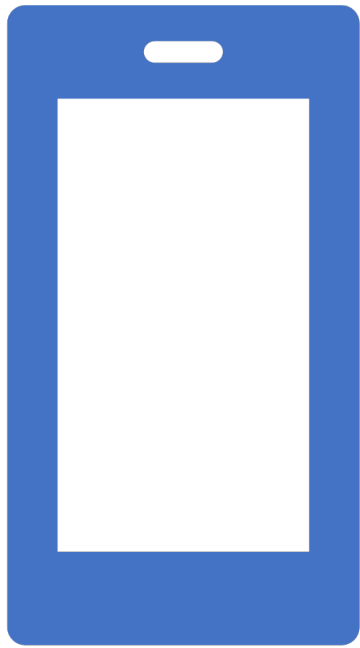
(Czaykowska-Higgins, 2009, p. 24)

Community-Based Research Cont.

Leonard and Haynes (2010) proposes a model of **true collaborative fieldwork**

Figure 2. A model of truly collaborative fieldwork (Leonard & Haynes, 2010, p.288)

Benefits of the Digital Verb Conjugator

Benefits of the digital verb conjugator

- Recording each of these combinations in a textbook or a dictionary would be very not practical if not impossible (approximately 20 years (A. Kazantseva, p.c., 2022))

- 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

© Matthew Holst / Press-Citizen

"101 Mohawk Verbs"?!?

16
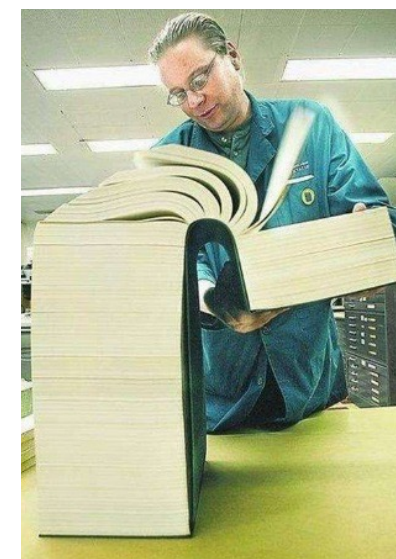
Technological approach

Why not machine learning?

- The training data for this problem simply doesn't exist.
  - This is a traditionally oral language and there is limited digital text available.
  - The verb forms attested in text are only a tiny slice of the tens or hundreds of thousands of possible forms.
- In this situation, neural NLG is just going to fabricate.
  - We want students learning *Oneida,* not a new language invented by an LM!

Also a language sovereignty issue!

- Part of our pledge is that the Oneida community remains in control of their language.

- Teachers get final say over what the answers are…

    - … and we can't guarantee that with a neural model.

    - They'd be effectively giving up some of their decision making capability to a black-box model.

    - With old-fashioned rule-based NLG, making changes to the output is straightforward; with neural NLG it's a research program (cf. Reiter, 2021).

So just make a handwritten FST, right?

- I approached the Kawennón:nis team to replicate their approach, but they brought up another issue.
  - As the project grew beyond a proof-of-concept, the code expanded in complexity until, realistically, only the original author could read/understand it.
  - Collaborators expressed concern about their inability to maintain and change their code if the original author moves on.
- This is another sovereignty issue!
  - Are they dependent forever on a third party if they ever need it to change?

XFST/LEXC code can become hard-to-read, especially when describing very morphologically complex languages.

An accessibility-focused DSL

- They suggested I use *Gramble* (Littell et al. 2024) instead, a XFST/LEXC-like language then in development.
  - Tabular syntax designed to be easy for non- and beginner programmers to read and change.
  - Live group-programming in a shared spreadsheet.
  - Easier to involve subject-matter experts (teachers, linguists, etc.) in grammar development
    - And even take over the project if/when the original programmer moves on.

Root =	text	gloss	
	kan	walk	
	pala	jump	
	ikar	climb	

Stem =	embed	text	person/gloss
	Root	ta	[1subj]
	Root	sa	[2subj]
	Root		[3subj]

replace text:	from	to	context
	t	d	(r|n)_

test:	text	gloss	
	ikarda	climb[1subj]	
	ikarta	climb[1subj]	

Gramble is a tabular language, whose canonical expression is a grid of cells.

Building the Digital Oneida Verb Conjugator

Gramble development

Data from publicly available resources, i.e.

- Twatati (2017)
- Michelson & Doxtator (2002)
- Michelson et al. (2016)
- Michelson (1988)
- Lounsbury (1976)

ACTV_hab=	root/text	root_uf	English/root_translation	colour	rootclass	Aspect_class	% comments
	yat	yat	_name(s)	purple	C	B2	
	kalatu·ni	kalatuni	_tell(s)_a_story_to	purple	C	A1	
	kalota·ni	kalotani	_give(s)_credit_to	purple	C	A1	
	kʌhlu	kʌhlu	_belittle(s)	purple	C	A1	%%
	kʌhla'slu·ni	kʌhla'sluni	_mistreat(s)/abuse(s)	purple	C	A1	se two verbs have same meaning
	kʌhlu·ni	kʌhluni	_mistreat(s)/abuse(s)	purple	C	A1	
	khunyʌ·ni	khunyʌni	_cook(s)_for	purple	C	A1	
	khwa	khwa	_take(s)_something_away_from	purple	C	D2	
	khwanu·tu	khwanutu	_feed(s)	purple	C	A1	
	kúhlek	kuhlek	_fist-beat(s)	purple	C	E2	%%
	kwaht	kwaht	_invite(s)	purple	C	E1	%%
	la·ni	lani	_serve(s)	purple	C	A1	
	lhalatstʌ·ni	lhalatstʌni	_promise(s)_someone_(nonbinding)	purple	C	A1	
	lihowanaht	lihowanaht	_honour(s)	purple	C	E1	%%
	lihunyʌ·ni	lihunyʌni	_teach(s)	purple	C	A1	
	lihwalho	lihwalho	_accuse(s)	purple	C	E2	
	slʌhto·li	slʌhtoli	_keep(s)_awake	purple	C	A1	
	naktota·ni	naktotani	ive(s)_permission/opportunity/time/spa	purple	C	A1	
	na'ku·ni	na'kuni	_make(s)_mad	purple	C	A1	
	na'tu·ni	na'tuni	_show(s)_something_to	purple	C	A1	%%
	nʌskwayʌ	nʌskwayʌ	_beg(s)	purple	C	B1	
	nhlalho	nhlalho	_give(s)_germs_to	purple	C	E2	
	noluhkw	noluhkw	_love(s)	purple	C	E1	%%
	thala·ni	thalani	_advise(s)	purple	C	A1	
	tsihalatenyá't	tsihalatenya't	_spin(s)	purple	C	E1	%%
	tsi'nyuhklokew	tsi'nyuhklokew	_wipe(s)_snot_off_of	purple	C	D3	

Figure 3. Entries of verb roots

	purple_PPR=	table:	subject_base/subject	English/subject_translation	text	embed	object_base/object	English/object_translation
3	%%		1SG	I>	XXX	C	1SG	>I
4	%%		1SG	I>	XXX	C	1DUEXCL	>someone_and_I
5	%%		1SG	I>	XXX	C	1DUINCL	>you_and_I
6	%%		1SG	I>	XXX	C	1PLEXCL	>they_and_I
7	%%		1SG	I>	XXX	C	1PLINCL	>you_two/all_and_I
8			1SG	I>	ku	C	2SG	>you
9			1SG	I>	kni	C	2DU	>you_two
10			1SG	I>	kwa	C	2PL	>you_all
11			1SG	I>	li	C	3SGM	>he
12	%%		1SG	I>	XXX	C	3SGFZ	>she/it
13			1SG	I>	khe	C	3FI	>she/they
14			1SG	I>	khe	C	3DUM	>they_two_(Ms)
15			1SG	I>	khe	C	3PLM	>they(Ms)
16			1SG	I>	khe	C	3DUFZ	>they_two_(Fs)
17			1SG	I>	khe	C	3PLFZ	>they_(Fs)
18	%%		1DUEXCL	someone_and_I>	XXX	C	1SG	>I
19	%%		1DUEXCL	someone_and_I>	XXX	C	1DUEXCL	>someone_and_I
20	%%		1DUEXCL	someone_and_I>	XXX	C	1DUINCL	>you_and_I
21	%%		1DUEXCL	someone_and_I>	XXX	C	1PLEXCL	>they_and_I
22	%%		1DUEXCL	someone_and_I>	XXX	C	1PLINCL	>you_two/all_and_I
23			1DUEXCL	someone_and_I>	kni	C	2SG	>you
24			1DUEXCL	someone_and_I>	kwa	C	2DU	>you_two
25			1DUEXCL	someone_and_I>	kwa	C	2PL	>you_all
26			1DUEXCL	someone_and_I>	shakni	C	3SGM	>he
27	%%		1DUEXCL	someone_and_I>	XXX	C	3SGFZ	>she/it
28			1DUEXCL	someone_and_I>	yakhi	C	3FI	>she/they
29			1DUEXCL	someone_and_I>	yakhi	C	3DUM	>they_two_(Ms)
30			1DUEXCL	someone_and_I>	yakhi	C	3PLM	>they(Ms)

Right panel fields:

subject_translation, rootclass, Aspect_class, root, root_uf, root_translation, colour, Habitual_Suffix, object, object_translation, subj_person, subj_number, subj_gender, subj_inclusivity, obj_person, obj_number, obj_gender, obj_inclusivity, text

Buttons: Download source | Generate & download [CSV] | Generate to new sheet | Sample to new sheet | Sample here

Figure 4. Entries of pronominal prefixes in the transitive sheet

replace text:	from	to	context
	7	e	kh|sh|sk|sl|st|th|tsh|tsy|'|kw)
	7		
	H	h	ı|á|é|í|ú|ó|ʌ|t|k|n|l|s|y|w|'|·)_
	H		

Figure 5. Rules that create variations for the pronominal prefix

User Interface Pilot Version

- https://yanfeilu.github.io/Oneida_Verb_Conjugator/

Evaluations and Results

Evaluation

- Transitive pronominal prefixes only

- Active class and habitual aspect only

- 56 verb roots are included in the database (74 forms if duplications that reflect phonological variations are also counted)

- 8 replacement rules (4 before the attachment of the pronominal prefix and 4 after)

- The combination of transitive pronominal prefixes, active and transitive verb roots, and habitual aspect suffixes lead to the generation of 8475 unique forms of conjugated verbs

The Test Set

- Contains 100 forms of conjugated examples of entries of the *Oneida English/English Oneida Dictionary* (Michelson & Doxtator, 2002)
  - Obviously not large enough to draw conclusions about.
  - More a hedge against the possibility of the programmer "overfitting" to known data.
- Although this is not an ML project, we adopted a dev/test division.
  - We had noticed some past projects evaluating on the same dataset as they used for development, making it hard to estimate their future performance on new data.
- Stress is not marked for most of the verbs, so the evaluation disregards prosodic features and examines only the alignment of the segments

Results

- The accuracy rate is 93% for both the "dev" data and the "test" data
  - "Dev": 44 forms included, 41 are correct
  - Incorrect forms:
    - *lakenhlálhos* 'He keeps giving me his germs'
    - **shako'tanʌwʌhslályo* 'He keeps whipping her'
    - *kheyahta'nawʌ́sta'* 'I dress her up warmly'
    - Each of the three incorrect forms are caused by a different reason

  - "Test": 56 forms included, 52 are correct

Feedback from the
Speakers and Learners

Feedback #1

- The learners and speakers suggested that the verb conjugator should also include audio recordings

- 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
  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

Feedback #2

- Concerns about ownership of the verb conjugator
  - Reassurance that neither I nor the ILT team will 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 organization of the community should be the optimal owner of the verb conjugator remains to be decided

- At any point of this research, we will ensure all collaborations are built on a foundation of reciprocity

Feedback #3

- Questions about the ongoing expenses associated with the website's upkeep and its hosting platform

- Everything runs in the browser, eliminating the need to provision a server. Currently the interface is hosted for free using GitHub Pages

- If the future owner of the conjugator chooses to migrate the interface and/or data, changes can be achieved with manageable effort and expense

Feedback #4

- Learners and speakers 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 digital learning materials

- The participants trust that the verb conjugator will make significant contributions to the revitalization of the Oneida language

Future Steps

Next Steps

- Adding additional paradigms
- Distinguishing attested from unattested forms in the interface
- Native-speaker verification of unattested forms
- Integrating prosodic (e.g. stress) rules

Next Steps

- We have only completed a pilot version and intend to continue collaborating with the partners and further refine this project

- 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

- We plan to step back from the front lines of development while remain committed to providing support and assistance when required

Acknowledgement

Collaborators from the Twatati Oneida language committee: Nancy George, Ursula Doxtator, and Tania Granadillo

Members of the The Indigenous Language Technologies Project Team of the NRC

Dr. Karin Michelson

yawʌ'ko

Thank you

Reference

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

- Brzozowski, J. (1964). *Derivatives of Regular Expressions. Journal of the ACM, 11*(4), 481–494. https://doi.org/10.1145/321239.321249

- 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.

- Koenig, J.-P., & Michelson, K. (2015). Invariance in argument realization: The case of Iroquoian. *Language*, *91*(1), 1–47. https://doi.org/10.1353/lan.2015.0008

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

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

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

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

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