NRC·CNRC

NRC Publications Archive
Archives des publications du CNRC

Fitting a square peg into a round hole: creating a UniMorph dataset of Kanien'kéha verbs

Kazantseva, Anna; Martin, Akwiratékha; Michelson, Karin , Koenig, Jean-Pierre

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 Seventh Workshop on the Use of Computational Methods in the Study of Endangered Languages, pp. 39-51, 2024-03-21

NRC Publications Archive Record / Notice des Archives des publications du CNRC :
https://nrc-publications.canada.ca/eng/view/object/?id=1e2f1d4a-e29d-4cca-a8d1-ef6d8aaa1637
https://publications-cnrc.canada.ca/fra/voir/objet/?id=1e2f1d4a-e29d-4cca-a8d1-ef6d8aaa1637

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

Fitting a Square Peg into a Round Hole: Creating a UniMorph dataset of Kanien'kéha Verbs

Anna Kazantseva and **Akwiratékha Martin**
National Research Council Canada
[anna.kazantseva, akwiratekha.martin]@nrc-cnrc.gc.ca

Karin Michelson and **Jean-Pierre Koenig**
University at Buffalo
The State University of New York
[kmich,jpkoenig]@buffalo.edu

Abstract

This paper describes efforts to annotate a dataset of verbs in the Iroquoian language Kanien'kéha (a.k.a. Mohawk) using the UniMorph schema (Batsuren et al., 2022a). The dataset is based on the output of a symbolic model - a hand-built verb conjugator. Morphological constituents of each verb are automatically annotated with UniMorph tags. Overall the process was smooth but some central features of the language did not fall neatly into the schema which resulted in a large number of custom tags and a somewhat *ad hoc* mapping process. We think the same difficulties are likely to arise for other Iroquoian languages and perhaps other North American language families. This paper describes our decision making process with respect to Kanien'kéha and reports preliminary results of morphological induction experiments using the dataset.

1 Introduction

It is generally believed that building language technology for morphologically rich languages benefits from knowing about morphology. Other things held constant, providing morphological information as a part of an NLP pipeline is likely to help, *e.g.* (Vania et al., 2018; Dehouck and Denis, 2018; Hofmann et al., 2021; Park et al., 2021). While there is no clear-cut definition of what makes a language morphologically rich, usually it refers to languages where words are composed of many parts. It certainly applies to most language families in North America (e.g. Algonquian, Iroquoian, Eskimo-Aleut, *etc*.)

Computational models of morphology are important also because mastering morphology is crucial when learning a morphologically rich language. Methods, techniques and aids that help students master morphology are helpful in speeding up the learning process (Renard, 2022).

The work described in this paper is a small step in this direction for Kanien'kéha, a.k.a. the Mohawk language. It started as a collaboration between the National Research Council Canada (further NRC) and a Kanien'kéha immersion school for adults, Onkwawénna Kentyóhkwa. The teachers at the school noticed that students in the immersion classes struggled most with mastering verbal morphology and often created hand-made 'verb conjugators' as study aids. The role of the NRC was to help build an interactive verb conjugator that was aligned with the school's curriculum. To the best of our knowledge this was the first computational model of a subset of Kanien'kéha grammar. However, we were unable to use any data driven methods because of the extreme paucity of textual data in Kanien'kéha.

One of the side effects of creating a symbolic language model was the creation of a large dataset of verbs (1,419K conjugations), complete with inflectional information and morphological segmentation. We have mapped this dataset into the UniMorph framework.

The motivation for this paper is two-fold. Firstly, the paper serves as a reference document for a new dataset for morphological induction in Kanien'kéha. The paper documents the dataset itself as well as the arbitrary decisions made during the labelling process. The second goal of this paper is to illustrate that such references are necessary when creating datasets for low-resource languages, especially less documented ones. We demonstrate several paradigms in the language that cannot be adequately described using the UniMorph framework without defining a large number of custom labels (e.g. pronominal system, aspect system, transitivity etc.). In some cases, existing UniMorph dimensions and features seem acceptable but upon closer inspection applying them would be misleading. These remarks are not meant as a criticism of UniMorph, but rather as suggestions for future updates of the schema. This is especially so since the same properties are common not only to all

Proceedings of the The Seventh Workshop on the Use of Computational Methods in the Study of Endangered Languages, pages 39–51
March 21-22, 2024 ©2024 Association for Computational Linguistics

Iroquoian languages, but also to other language families in North America (e.g. Algonquian).

The main contribution of this work is the dataset[1]. The second contribution is preliminary results of morphological inflection experiments using this dataset. To the best of our knowledge, this is the first data-driven or corpus-based model of Kanien'kéha.

This paper is structured as follows. Section 2 places the work into existing research context. Section 3 provides a brief overview of the language. Section 4 gives an overview of UniMorph. Section 5 describes the initial data used for annotation. Section 6 is the main description of the new dataset and the decisions made. Section 7 briefly describes experiments and reports results. Sections 8 and 9 contain discussion and describe limitations of this work.

2 Related Work

Related work falls into two broad categories: linguistic and computational modelling of Kanien'kéha and research on computational models of morphology induction.

There is ample work in the field of Linguistics describing Kanien'kéha. Mithun (2000, 2005) provide thorough overviews of the language. Lounsbury (1953) describes the closest Iroquoian *sister* language - Oneida, and so do Michelson and Doxtator (2002). Beaty (1974) and Bonvillain (1973) are grammars of two dialects of Kanien'kéha. These resources describe the language as a system but we could not use them directly in computational modelling because of lack of both coverage and detail. A notable exception is Michelson (1983) which features a complete and detailed model of the stress system in Kanien'kéha; the symbolic model we have built is an implementation of this work.

Another type of descriptive work are educational materials: Maracle (2017); Martin (2023); Price et al. (2011). These are teaching textbooks and curriculum materials. As such they are complete, thorough and focus on the aspects of the language that are important for today's learners. We have used them extensively.

An important research hive for activity on computational models of morphology is the *Special Interest Group on Computational Morphology and Phonology* (*SIGMORPHON*). The annual shared task competitions (Nicolai et al., 2023) include morphological inflection. In 2023 the task was run on 26 languages across 9 language families. Systems that consistently perform better across languages are neural ones (e.g., Canby et al. 2020; Girrbach 2023). However for some languages a non-neural and rule-based systems designed specifically for those languages achieve best results (e.g., Kwak et al. 2023).

Within this context, our work is novel with respect to resources and research on computational modelling of Kanien'kéha. We have created the first large dataset of inflected verbs in Kanien'kéha that can be used in computational modelling. Our computational experiments at this point are basic - we use the SIGMORPHON neural character-level transformer baseline (Wu et al., 2021).

3 Kanien'kéha and Iroquoian Languages

Iroquoian languages are a group of approximately 17 historically documented languages situated in southeastern Canada (Ontario and Quebec) and northeastern US (New York State, but also in North Carolina and Oklahoma).

All spoken Iroquoian languages that still have first-language speakers (further *L1*) (Cherokee, Seneca, Cayuga, Onondaga, Oneida and Kanien'kéha) are endangered. Several others are either undergoing revitalization within communities or are considered sleeping languages. The majority of *L1* speakers are older than 75. However, in several communities a small number of new *L1* speakers are being raised by parents who are *L2* speakers.

Linguistically, Iroquoian languages can be divided into Southern Iroquoian and Northern Iroquoian branches. There is only one Southern Iroquoian language - Cherokee. The Northern branch of the Iroquoian language family contains all original Five Nations languages of the Haudenosaunee Confederacy. Many other Iroquoian languages are no longer spoken, with scant word lists available (e.g. Wyandot, Petun, Meherrin, Neutral, Wenro and Erie to name just a few) (Mithun, 2005).

Despite the current harsh linguistic reality, due to the effects of continued colonization and governments' efforts to linguistically and culturally destroy them, the language communities are very focused and interested in strengthening and re-establishing their ancestral languages. Many im-

[1]Due to the preference of the communities the dataset is not publicly available by default, however it is available upon request for research and educational purposes.

pressive and successful efforts are ongoing and evolving to fit their contemporary needs. The authors of the paper are only familiar with some of the communities and regret inevitable omissions. However, examples of thriving language schools are seen at Twatati Adult Oneida Immersion program for Oneida[2], Yawenda Project for Wendat[3] and Onkwawenna Kentyohkwa[4] and Kanien'kehá:ka Onkwawén:na Raotitióhkwa Language and Cultural Center for Kanien'kéha Ratiwennahní:rats Adult Language Immersion Program[5].

All Iroquoian languages are morphologically complex, with verbs being particularly elaborate. Verbs are composed of several parts, both prefixes and suffixes, as well as noun incorporation. Due to the languages' rich morphology, the linguistic practice of creating new words is deeply cultural, therefore restricting borrowing from other languages. Generally speaking, single-token verbs in an Iroquoian language correspond to simple clauses in English (however simpler verbs are possible too):

(1) *enhake'serehtakwatákwahse'*
en-hake-'sereht-a-kwatakw-a-hs-e'
will-he>me-car-link-fix-link-for-punctual
FUT-MSG>1SG-car-JR-fix-JR-BEN-PUNC

'he'll repair a car for me; he'll fix my car' (Kanien'kéha)

(2) *yusayenhohaya?ákhu?*
y-usa-ye-nhoh-a-ya?ak-hu-?
there-again.did-she-door-link-hit-many-punc
TRANSL-REP.FACT-FI.A-door-JR-hit-
-DISTR-PUNC

'she knocked on the door again' (Oneida)

Example 1 and Example 2 are two unremarkable Kanien'kéha and Oneida verbs (Michelson et al., 2016, p.51) that correspond to simple clauses in English.

Since this work only focuses on verbal morphology, we will only discuss that part of speech from here on in. A minimal verb structure consists of a pronominal prefix, a verb stem and an aspectual suffix, which can be null. A verb stem can be simple or have a noun incorporated, as in Example 1.

[2]https://www.facebook.com/people/Twatati/100057069505224
[3]https://languewendat.com/en/ and (Lukaniec, 2018)
[4]https://onkwawenna.info/
[5]https://www.korkahnawake.org/kanienkha-ratiwennahnrats

Classificatory dimensions
Aktionsart
Animacy
Argument Marking
Aspect
Case
Comparison
Definiteness
Deixis
Evidentiality
Finiteness
Gender
Information Structure
Interrogativity
Language-specific features
Mood
Number
Part of Speech
Person
Polarity
Politeness
Possession
Switch-reference
Tense
Valency
Voice

Table 1: Classificatory dimensions in UniMorph

Additionally, a verb can also have pre-pronominal prefixes and the verb stem can include one or more prefixes or suffixes; these convey inflectional and derivational meanings.

4 UniMorph

The UniMorph project (Sylak-Glassman, 2016; Batsuren et al., 2022b) has two main parts: an annotation schema and an extensive collection of inflection tables for 182 languages, several dozens of them for endangered languages of the world. It also contains several datasets for morpheme segmentation and for derivational morphology.

The original goal of the project was to develop a language-independent schema that could adequately describe inflectional morphological breakdowns of words in any language. Currently in its 4.0 version, the UniMorph schema has been extended and improved but its skeleton remains unchanged.

The UniMorph schema contains 25 dimensions listed in Table 1. Each dimension has a number of possible features. For example, the features

Stem	Inflected form	UniMorph tags	Lang.
mercify	mercifies	V;PRS;3;SG	EN
mercify	mercifying	V;V.PTCP;PRS	EN
abalienare	abalienò	V;IND;PST;3;SG;PFV	IT
abbacare	abbacai	V;IND;PST;1;SG;PFV	IT

Table 2: Examples of words annotated using the UniMorph schema.

for the dimension *Animacy* are Animate, Human, Inanimate, Non-human. When default features are insufficient the annotators can also create language specific tags.

Table 4 shows examples of words annotated using UniMorph.

Using the UniMorph schema is not the only option. We could have devised our own set of tags as in Hämäläinen et al. (2021). We decided to opt for standardization possibly at the expense of specificity. This decision is due to the practical nature of our objectives. As the overarching goal is improving language technology for Kanien'kéha, choosing a widely used standard (UniMorph) seems more practical and more accessible for future users than devising our own.

5 The dataset

Kawennón:nis is an interactive verb conjugator for Kanien'kéha available online or locally. Currently two version of the software exist: one for the Ohswé:ken dialect (Ohswekèn:'a) and another for the Kahnawà:ke dialect (Kahnawa'kéha). This paper describes the dataset for Kahnawa'kéha.

This tool was designed as a teaching aid for students of immersion programs of Kanien'kéha and closely follows the curriculum. It allows the user to select one or more verb stems, one or more sets of pronominal prefixes and one or more tenses, and then outputs the corresponding conjugations. The tool has been designed in close collaboration with Onkwawénna Kentyóhkwaimmersion school. Following a user study and several consultations with teachers and students, a local artist was employed to design a culturally relevant interface.

The tool contains 662 verb stems and more are being added. It allows the user to choose one or more of the 12 available tense/aspect options, as well as apply negation and repetition. It does not contain derivational morphology, although we hope to add contained subsets in the future.

The complete output of the tool corresponds to inflectional tables for the 662 verbs within the modeled paradigms - 1,418,939 inflected forms. However because the dataset contains inflectional tables as opposed to intelligently created samples, there is a lot of redundancy (the size grows exponentially with the number of paradigms modeled).

Stress in Kanien'kéha is quite complex and is a major source of irregularities. We make available both the stressed and the unstressed versions on the dataset.

Extensive work has been done to ensure the quality of the dataset but as any computational model, it contains errors. 244 of the 662 inflection table files have been manually checked by an advanced L2 speaker who is an experienced teacher and linguist. As was mentioned in Section 3 there are very few L1 speakers of Kanien'kéha; what is even more important is that their time is better spent than checking conjugation tables. We realize that thorough evaluation is a weakness of this work but it is an unfortunate consequence of the capacity bottleneck.

6 UniMorph and Kanien'kéha

Our dataset contains only active verbs (as opposed to stative) and captures only foundational morphological paradigms. Therefore, only verb-related parts of the UniMorph schema are relevant to us. In this section we describe our efforts to align the properties of the language with the UniMorph schema. The process of mapping was not straightforward and arguably alternative choices could have been made. Yet, despite some difficulties we were able to automatically label our existing dataset, take advantage of the existing systems and train a model for morphological inflection in Kanien'kéha. Preliminary results are available in Section 7.

6.1 Valency and Voice-like features

Depending on the structure of their semantic arguments, there are three types of verbs in Kanien'kéha: two intransitive types, and one transitive. Intransitive verbs can be divided into two classes depending on the category of bound pronominal prefixes they take. Two types of pronominal prefixes are possible with intransitive verbs: *agent* prefixes and *patient* ones. The distribution of the prefixes typically has to do with the degree of control the actor

has over the event.

(3) *ie'níkhons*
ie-'nikhon-hs
she/someone/they-sew-habitual
FI.A-sew-HAB

'she/they/someone sews or is sewing'

In Example 3 the actor (*she/someone/they*) is in control and the active pronominal prefix *ie-* (*she*) is used. If control is lacking or the actor is being acted upon, patient pronominal prefixes are used:

(4) *saho'nikónhrhen'*
s-wa-ro-'nikonhrhen-'
again-did-he-forget-punctual
REP-FACT-MSGP-forget-PUNC

'he forgot (again)'

In Example 4 the actor has less control over the event, hence the patient pronominal prefix *ro-* (*he/him*) is used. While the degree of control largely determines pronominal prefix preferences of a verb, it is not always the case and students must learn them for each verb. For instance, the verb *yo'ten* (*to work*) always takes patients pronominal prefixes:

(5) *enionkeniió'ten'*
en-ionkeni-io'ten-'
will-we-work-punctual
FUT-1DUP-work-PUNC

'we will work'

Also, patient pronominal prefixes are always used in certain tense/aspect combinations that emphasize the end result of an action:

(6) *wakatórion*
wak-atori-on
I-drive-stative
1SGP-drive-STAT

'I have driven (emphasis on the result)'

For transitive verbs when both participants in the event are animate, a separate set of pronominal prefixes is used to encode the relation between the participants, as in the following example:

(7) *taiethi'nikonhrakénnion*
t-a-iethi-'nikonhr-a-kenni-on
two-should-we>them-mind-link-challenge-stative
DUP-OPT-1INCL.NS>3NS-mind-JR-compete-STAT

'We should have convinced them.'

In Example 7 the transitive pronominal prefix *yethi-* is used, meaning *you-and-I/we-to-her/her/them*.

While there is a semantic distinction (based on the degree of control and the relationship between agents and patients), pronominal prefix preferences are sometimes lexicalized in intransitive cases. The three types of verbs are learned in the first lessons of Kanien'kéha and students need to memorize the type of pronominal prefixes each verb takes (some verbs can participate in constructions of more than one type).

These paradigms roughly correspond to two of the UniMorph dimensions: Valency (the transitive/intransitive distinction) and Voice (Active/Passive) distinction. However, we decided against using these default categories. The first reason is that even intransitive constructions such as those in Example 3 and Example 4 can be used with semantically transitive verbs - in those cases the pronominal prefix means *actor-to-it*. For instance, the verb *ie'níkhons* can mean *'She is sewing something'* (where the *'something'* is understood). So the distinction is not strictly in the number of semantic arguments. Secondly, the Voice dimension of UniMorph and Active/Passive distinction does not correspond to semantic differences of agent and patient pronominal prefixes and corresponding constructions. Voice alternations mark situations when the relationship between a verb and its core nominal arguments is altered. The distinction in Kanien'kéha is different; there is no true voice alternation in the language.

6.2 Pronominal prefix features

Verbs in Kanien'kéha require a bound pronoun to be grammatical. Bound pronouns are often referred to as *pronominal prefixes* and that is the terminology we use throughout this paper. The pronominal prefix signifies a relationship between an agent and a patient (e.g. *he-to-it*, *you and I-to-those two*, *it-to-me*). The pronominal system in Kanien'kéha is very complex and elaborate.

The Kanien'kéha pronominal system distinguishes person (*1st, 2nd, 3rd*) and number (singular, dual, plural). *1st* person dual and plural pronominal prefixes also mark for inclusivity of the listener (e.g. *teni-* (*you and I*) vs. *iakeni-* (*someone and I*)). There are three gender choices: masculine, feminine/indefinite (the indefinite pronominal prefix is identical to feminine across the pronominal prefix system) and feminine/zoic (referring to some female persons and animals).

Annotating person, number, gender and inclusivity categories within the UniMorph framework is straight-forward. The only thing worth noting is the distinction between the transitive and the quasi-intransitive verbs. Recall that intransitive verbs in Kanien'kéha can 1) truly take one semantic argument as in *teharáhtats* (*'he runs'*) or 2) they can denote a relationship between an animate and inanimate entity as in *wahahní:non'* (*'he bought (it, something)'*). For intransitive verbs we only annotate the agent (or the patient), but we do not explicitly annotate the implicit participant (*it*). For transitive pronominal prefixes we annotate both the agent and the patient.

6.3 Tense-related features

When looking at the verb conjugator Kawennón:nis and at the instructional texts for Kanien'kéha immersion courses we often see the term *tense*. Yet, the notion of tense in Kanien'kéha is quite different and the terminology is influenced by the fact that most teachers and students are L1 speakers of English rather than by the intrinsic semantics of Kanien'kéha.

Tense refers to the relationship between the time of utterance (TU) and topic time (TT) with other refinements possible (Reichenbach, 1947; Klein, 1994). However, in Kanien'kéha the meaning of tense is intertwined with the meaning of mood-related categories of *realis* and *irrealis*. So what we refer to as the past tense, in Kanien'kéha is closer to the marker of something having happened for sure; present time - happening at the time of utterance; future tenses refer to likely events in the future and optative constructions - to possible future events.

(8) *ie'níkhons*
ie-'nikhon-hs
she/someone/they-sew-habitual
FI.A-sew-HAB

'she sews it (either habitually or right now)'

(9) *enie'níkhon'*
en-ie-'nikhon-'
will-she/someone/they-sew-punctual
FUT-FI.A-sew-PUNC

'She will sew it (definitely)'

(10) *wa'e'níkhon'*
wa'-ie-nikhon-'
did-she/someone/they-sew-punctual
FACT-FI.A-sew-PUNC

'She sewed, she did sew it (it is a fact)'

(11) *aie'níkhon'*
a-ie '-nikhon-'
should-she/someone/they-sew-punctual
OPT-FI.A-sew-PUNC

'She should, might, or ought to sew it'

Our dataset contains two tenses that we label as past: 1) punctual factual and 2) habitual with former past.

Verbs with explicit markers of 'future' are labelled as future tense *(FUT)* and those with former past suffix *-kwe'* as past tense *(PST)*. We do not explicitly label what most students think of the present tense.

6.4 Aspect-related features

The situation with aspect is no less complicated. The notion of aspect is based on the relationship between Time of Situation (TSit) and Topic Time (TT) (Reichenbach, 1947; Klein, 1994). The UniMorph schema also defines aspect as the relationship between the time for which a claim is made (TT) and the time for which a situation actually held true (TSit) (Sylak-Glassman, 2016) (page 13).

Linguistic literature on Kanien'kéha (Beaty, 1974; Bonvillain, 1973; Price et al., 2011; Martin, 2023) varies somewhat in their labelling and the number of aspects. Recent work (Price et al., 2011; Martin, 2023) agrees on distinguishing three

aspects: Imperfective (a.k.a. Habitual), Punctual (a.k.a. Perfective) and Stative, plus Imperative (which is marked by the absence of any aspectual information).

These categories are neither orthogonal nor parallel to the UniMorph labels which are *Imperfective, Perfective, Perfect, Progressive, Prospective, Iterative, Habitual*.

In Kanien'kéha, the Habitual aspect can denote three possible cases 1) Habitual occupation or profession 2) Daily recurring activity and 3) An event occurring at the moment of speech.

For example the verb *ie'níkhons* in Example 8 can be translated as 1) She is a seamstress 2) She sews (regularly) or 3) She is sewing right now.

We decided to label the Habitual aspect in Kanien'kéha using the UniMorph *Imperfective* label (*IPFV*), e.g. Example 8.

Punctual aspect can be combined with factual (Example 10), future (Example 9) or optative constructions (Example 11). It denotes actions that are viewed as complete events and approximately corresponds to *Perfective* aspect in UniMorph. We label constructions in punctual aspect as perfective using the UniMorph schema (*PFV*). Punctual factual constructions usually are translated as past tense, as in the Example 10 yet it is not truly a tense. Factual *wa'-* is a mood marker and the emphasis is on factuality and certainty, not tense (however, it is commonly translated as Simple Past into English).

Progressive meaning in Kanien'kéha is sometimes expressed with the habitual aspect, as in *ie'níkhons*: (*'she is sewing'*) in Example 8 and sometimes with the stative aspect, as in *wakatshenón:ni* (*'I am happy'*) in Example 12. There seems to be a correlation with the notions of accomplishment versus activity, but this varies so much that the distribution has to be learned for each verb. For Stative constructions, we define a language-specific label. (The stative also can convey the equivalent of the English perfect, as in Example 13, *tewaktà:on' 'I have stopped'*. We use the UniMorph label *Perfect* (*PFV*) for such cases.)

(12) *wakatshennón:ni*
wak-atshennonni-ø
I-happy-stative
1SGP-happy-STAT

'I am happy'

(13) *tewaktà:'on*
te-wak-t-a-'-on
two-I-stand-link-become-stative
DUP-1SGP-stand–JR-INCH-STAT

'I have stood up, I have stopped'

Two types of constructions that do not fit into the UniMorph dimensions are Perfect Progressive and Habitual Continuative constructions.

(14) *iako'nikhóntie'*
iako-'nikhón-ø-tie'
she-sew-stative-progressive
FI.P-sew-STAT-PROG

'She is sewing it (while moving along in space and or in time)'

(15) *enie'nikhónhseke'*
en-ie-'nikhón-hs-eke'
will-she-sew-habitual-continuative
FUT-FI.P-sew-HAB-CONT

'She will keep on sewing it'

The semantics of the Perfect Progressive in Kanien'kéha does not fit neatly into the aspectual hierarchy. For Perfect Progressive the semantic component of motion is as important as that of continuity, see Example 14. We define a language-specific feature for this type of constructions. Habitual continuative tenses (optative and future) emphasize events with duration, as in Example 15. Yet the markers seem to be formal elements that convey the usual semantics of habitual. We do not add a feature for these constructions.

6.5 Mood-related features

We annotate three possible feature values for the *Mood* dimension: *Imperative* (*IMP*) for commands, *Irrealis* (*IRR*) for optative constructions, and *Realis* (*REAL*) for past tense and factual constructions.

6.6 Finiteness features

In our annotation we annotate Perfect and Perfect Optative tenses as *Finite* (*FIN*). We do not explicitly mark *Nonfinite* constructions.

6.7 Deixis features

Deixis is a linguistic mechanism for referring to a location, entity or time within a given context. For

instance, the use of words such as *this, that, here, then, etc.* are examples of deixis.

Kanien'kéha verbs can take on non-modal pre-pronominal prefixes that have deictic properties:

(16) *entkontáweia'te'*
en-t-kon-ataweia't-e'
will-here-they-enter-punctual
FUT-CIS-FZPLA-enter-PUNC

'they will come in, they will enter (here)'

(17) *ienkontáweia'te'*
i-en-kon-ataweia't-e'
there-will-they-enter-punctual
TRANS-FUT-FZPLA-enter-PUNC

'they will go in, they will enter (there)'

We define two language specific features to address such cislocative (Example 16) and translocative (Example 17) constructions.

The same morphological slot can be occupied by a number of other semantic prefixes that are not diectic in nature. Nevertheless, we list them in this section. The *TE* prefix is a dualic prefix that denotes various pair-wise relationships. The *NI* prefix is a partitive prefix that has several meanings: quantitative, intensity or location. There is also the *S* prefix denoting repetition. We defined three language specific tags to denote the meaning of these prefixes.

(18) *tesewakwáthon*
te-sewa-kwath-on
two-you-hem-stative
DUP-2PLP-hem-STAT

'you have hemmed it'

(19) *tho* *naieseniié:renke'*
tho n-aie-seni-ier-en-ke'
there partitive-should-you-do-st
partical PART-OPT-2DUP-do-STAT-CONT

'you should have done that'

Dataset	Accuracy	Edit distance
With stress	64.93	1.28
Unstresed	75.36	0.83

Table 3: Experimental results

(20) *ia'tesewakerihwaientà:se'*
ia'-te-se-wake-rihw-a-ient-a-'s-e'
there-two-again-I-issue-link
settle.on.ground-link-for-punctual
TRANS-DUP-REP-1SG-issue-
JR-put.down-JR-BEN-STAT

'I have decided again'

7 Experiments

For our experiments we apply the neural character-level transformer (Wu et al., 2021) that is used as a competitive baseline in SIGMORPHON competitions (Nicolai et al., 2023). We use the high-resource setting and the only parameter we change from default settings is the batch size, which we set to 1000[6].

The final version of the dataset contains 1,418,893 inflected verb forms. We split them into training, development and test sets in a $0.7 : 0.1 : 0.2$ ratio. The splits are done at the level of individual verbs: the same verb does not appear in more than one split (and consequently none of the conjugated examples are overlapping either). The exception to this statement are verbs that can behave as both transitive and intransitive ones: in these cases while the stem is the same, the morphological preferences are different. Because of this we do not control that such verbs do not appear in training/development sets and test sets simultaneously.

The stress system is a major source of exceptions. In addition to placement of stress, it also determines variations in length and tone and whether some characters are omitted. Because of these challenges we create a second dataset without stress. To produce the dataset, we use the output of the symbolic model before stress rules are applied.

The results are shown in Table 3. As expected, the task of producing conjugations without stress is much easier: the system achieves 75.36% accuracy vs. 64.93% on the stressed dataset.

[6]This follows advice from Wu et al. (2021) who find that batch size plays a critical size for transformer-based models of morphology.

Preliminary error analysis suggests that indeed stress-related errors are common. However the biggest source of mistakes seems to be changes in verb stems related to aspect (habitual, punctual, stative and imperatives). This is supported by opinions of the speakers that aspectual endings are largely lexicalized with many exceptions from general loose rules. In the symbolic system used as the original source of data, four forms are given as input for each verb stem. The neural system learned how to generalize those forms, albeit imperfectly.

We hope to address these shortcomings in future work. Since stress in Kanien'kéha is determined from the end of the word, in right-to-left fashion, we expect that applying a right-to-left system such as that of Canby et al. (2020) may help. Also, we hope that learning generic rules about orthography in Kanien'kéha from an existing corpus may improve both stress and aspectual-class related errors.

8 Conclusions

We have described the first dataset for morphological induction for the Iroquoian language Kanien'kéha. Due to community preferences, the dataset is not publicly available by default but is available upon request for research and educational purposes.

While describing the dataset, we have demonstrated that the process of mapping ready morphologically segmented data into UniMorph is neither trivial nor always straight-forward. Some of the problematic categories, common to other Iroquoian languages, are tense, aspect, voice and mood. We have also reported first results of morphological induction experiments on this dataset.

In the future, we have practical and research directions to consider. We hope to use the results of experiments to speed up the circular process of improving the symbolic model, extending this dataset and hopefully eventually exceeding the precision of the symbolic model.

We also are exploring ways to improve performance on this dataset. One such avenue was mentioned in Section 7: using a learning model that considers both left-to-right and right-to-left input directions. We also would like to look into creating similar resources for related languages for which symbolic models already exist, *e.g.* Oneida (Lu, 2023) and Cherokee[7].

[7] https://www.yourgrandmotherscherokee.com

9 Limitations

This work is one small step in the direction of applying data-driven language technology to Kanien'kéha. As such, its limitations are plentiful.

The most obvious one is that the dataset covers only a subset of the language: only active verbs, and only foundational verbal paradigms. Despite the fact that verbs are the most complex and common part-of-speech in Kanien'kéha, the performance on this dataset may not generalize sufficiently well.

The second limitation is the precision of the dataset. The source of the dataset is a symbolic model built by hand and checked by hand. We have done our best (and continue to do so) to gradually check the correctness a large part of the conjugated forms (244 verbs have been manually checked). Yet, without a doubt some errors remain.

Another limitation is generalizing to other languages. The UniMorph repository contain a dataset for one more Iroquoian language - Seneca (Pimentel et al., 2021); it is even more limited in scope than ours. We would like to create datasets for related languages, but that is only possible for cases where there already exist high quality resources.

The fourth, perhaps most crucial limitation is in the applications of the dataset and of the experimental results. Kanien'kéha is an endangered language spoken by several hundred people across several communities. In this context it is crucial to check every step for whether a given technology or resource helps or hurts the vitality of the language, or has the potential to do so. It is not immediately clear how to use this resource especially given concerns about data governance and sovereignty. We intend to use our results to identify the most difficult verbs and create a feedback loop. This may or may not be helpful. We hope others may come up with better use cases.

Ethics Statement

When working with endangered languages ethical concerns are paramount. All Indigenous languages spoken in Canada have a history of language suppression, expropriation and, at times, misuse. In this historical context unhurried discussions with the language communities, genuine partnership in creating resources and software and informed consent are the bare minimum. It is not difficult to see that this requirement is likely to slow down the technical side. It is because of these concerns that

we decided not to release the dataset by default but to make it available upon request.

Another ethical concern is how creation of a resource can affect the language - especially in situations where there are very few or no digital resources. For Kanien'kéha the writing system is fairly recent and the orthography is not always consistent. Creation of a resource can influence the standards of spelling - sometimes incorrectly so. This is especially dangerous in situations where few people are confident enough in their spelling to point out a mistake.

Acknowledgements

The authors of the paper would like to acknowledge extensive help and contributions of the personnel and of the students of Onkwawénna Kentyóhkwa immersion school, especially Owennatékha Brian Maracle, Karakwenhawi Zoe Hopkins, Rohahí:yo Jordan Brant, Jody Maracle, Ronkwe'tiyóhstha Josiah Maracle and Maura Abrams. We are grateful to Aidan Pine for his work on the server-side and GUI of Kawennón:nis and for multiple fruitful discussions. We also thank our colleagues Roland Kuhn and Patrick Littell for their help with this project.

References

Khuyagbaatar Batsuren, Omer Goldman, Salam Khalifa, Nizar Habash, Witold Kieraś, Gábor Bella, Brian Leonard, Garrett Nicolai, Kyle Gorman, Yustinus Ghanggo Ate, Maria Ryskina, Sabrina Mielke, Elena Budianskaya, Charbel El-Khaissi, Tiago Pimentel, Michael Gasser, William Abbott Lane, Mohit Raj, Matt Coler, Jaime Rafael Montoya Samame, Delio Siticonatzi Camaiteri, Esaú Zumaeta Rojas, Didier López Francis, Arturo Oncevay, Juan López Bautista, Gema Celeste Silva Villegas, Lucas Torroba Hennigen, Adam Ek, David Guriel, Peter Dirix, Jean-Philippe Bernardy, Andrey Scherbakov, Aziyana Bayyr-ool, Antonios Anastasopoulos, Roberto Zariquiey, Karina Sheifer, Sofya Ganieva, Hilaria Cruz, Ritván Karahóǧa, Stella Markantonatou, George Pavlidis, Matvey Plugaryov, Elena Klyachko, Ali Salehi, Candy Angulo, Jatayu Baxi, Andrew Krizhanovsky, Natalia Krizhanovskaya, Elizabeth Salesky, Clara Vania, Sardana Ivanova, Jennifer White, Rowan Hall Maudslay, Josef Valvoda, Ran Zmigrod, Paula Czarnowska, Irene Nikkarinen, Aelita Salchak, Brijesh Bhatt, Christopher Straughn, Zoey Liu, Jonathan North Washington, Yuval Pinter, Duygu Ataman, Marcin Wolinski, Totok Suhardijanto, Anna Yablonskaya, Niklas Stoehr, Hossep Dolatian, Zahroh Nuriah, Shyam Ratan, Francis M. Tyers, Edoardo M. Ponti, Grant Aiton, Aryaman Arora, Richard J. Hatcher, Ritesh Kumar, Jeremiah Young, Daria Rodionova, Anastasia Yemelina, Taras Andrushko, Igor Marchenko, Polina Mashkovtseva, Alexandra Serova, Emily Prud'hommeaux, Maria Nepomniashchaya, Fausto Giunchiglia, Eleanor Chodroff, Mans Hulden, Miikka Silfverberg, Arya D. McCarthy, David Yarowsky, Ryan Cotterell, Reut Tsarfaty, and Ekaterina Vylomova. 2022a. UniMorph 4.0: Universal Morphology. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 840–855, Marseille, France. European Language Resources Association.

Khuyagbaatar Batsuren, Omer Goldman, Salam Khalifa, Nizar Habash, Witold Kieraś, Gábor Bella, Brian Leonard, Garrett Nicolai, Kyle Gorman, Yustinus Ghanggo Ate, Maria Ryskina, Sabrina Mielke, Elena Budianskaya, Charbel El-Khaissi, Tiago Pimentel, Michael Gasser, William Abbott Lane, Mohit Raj, Matt Coler, Jaime Rafael Montoya Samame, Delio Siticonatzi Camaiteri, Esaú Zumaeta Rojas, Didier López Francis, Arturo Oncevay, Juan López Bautista, Gema Celeste Silva Villegas, Lucas Torroba Hennigen, Adam Ek, David Guriel, Peter Dirix, Jean-Philippe Bernardy, Andrey Scherbakov, Aziyana Bayyr-ool, Antonios Anastasopoulos, Roberto Zariquiey, Karina Sheifer, Sofya Ganieva, Hilaria Cruz, Ritván Karahóǧa, Stella Markantonatou, George Pavlidis, Matvey Plugaryov, Elena Klyachko, Ali Salehi, Candy Angulo, Jatayu Baxi, Andrew Krizhanovsky, Natalia Krizhanovskaya, Elizabeth Salesky, Clara Vania, Sardana Ivanova, Jennifer White, Rowan Hall Maudslay, Josef Valvoda, Ran Zmigrod, Paula Czarnowska, Irene Nikkarinen, Aelita Salchak, Brijesh Bhatt, Christopher Straughn, Zoey Liu, Jonathan North Washington, Yuval Pinter, Duygu Ataman, Marcin Wolinski, Totok Suhardijanto, Anna Yablonskaya, Niklas Stoehr, Hossep Dolatian, Zahroh Nuriah, Shyam Ratan, Francis M. Tyers, Edoardo M. Ponti, Grant Aiton, Aryaman Arora, Richard J. Hatcher, Ritesh Kumar, Jeremiah Young, Daria Rodionova, Anastasia Yemelina, Taras Andrushko, Igor Marchenko, Polina Mashkovtseva, Alexandra Serova, Emily Prud'hommeaux, Maria Nepomniashchaya, Fausto Giunchiglia, Eleanor Chodroff, Mans Hulden, Miikka Silfverberg, Arya D. McCarthy, David Yarowsky, Ryan Cotterell, Reut Tsarfaty, and Ekaterina Vylomova. 2022b. UniMorph 4.0: Universal Morphology. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 840–855, Marseille, France. European Language Resources Association.

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

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

Marc E. Canby, Aidana Karipbayeva, Bryan Lunt, Sahand Mozaffari, Charlotte Yoder, and Julia Hock-

enmaier. 2020. University of illinois submission to the SIGMORPHON 2020 shared task 0: Typologically diverse morphological inflection. In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology, SIGMORPHON 2020, Online, July 10, 2020*, pages 137–145. Association for Computational Linguistics.

Mathieu Dehouck and Pascal Denis. 2018. A framework for understanding the role of morphology in Universal Dependency parsing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2864–2870, Brussels, Belgium. Association for Computational Linguistics.

Leander Girrbach. 2023. Tü-CL at SIGMORPHON 2023: Straight-through gradient estimation for hard attention. In *Proceedings of the 20th SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 171–185, Toronto, Canada. Association for Computational Linguistics.

Mika Hämäläinen, Niko Partanen, Jack Rueter, and Khalid Alnajjar. 2021. Neural morphology dataset and models for multiple languages, from the large to the endangered. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 166–177, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Valentin Hofmann, Janet Pierrehumbert, and Hinrich Schütze. 2021. Superbizarre is not superb: Derivational morphology improves BERT's interpretation of complex words. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3594–3608, Online. Association for Computational Linguistics.

Wolfgang Klein. 1994. *Time in Language*. Routeledge.

Alice Kwak, Michael Hammond, and Cheyenne Wing. 2023. Morphological reinflection with weighted finite-state transducers. In *Proceedings of the 20th SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 132–137, Toronto, Canada. Association for Computational Linguistics.

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

Yanfei Lu. 2023. Empowering the oneida language revitalization: Development of an oneida verb conjugator.

Megan Lukaniec. 2018. *The elaboration of verbal structure: Wendat (Huron) verb morphology*. Ph.D. thesis, University of California, Santa Barbara.

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

Akwiratékha' Martin. 2023. *Tekawennahsonterónnion: Kanien'kéha Morphology*. Kanien'kehá:ka Onkwawén:na Raotitióhkwa Language and Cultural Center.

Karin Michelson and Mercy Doxtator. 2002. *Oneida-English/English Oneida Dictionary*. University of Toronto Press.

Karin Michelson, Norma Kennedy, and Mercy A. Doxtator. 2016. *Glimpses of Oneida Life*. University of Toronto Press.

Karin Eva Michelson. 1983. *A Comparative Study of Accent in the Five Nations Iroquoian Languages (Mohawk, Oneida, Onondaga, Cayuga, Seneca)*. Ph.D. thesis, Harvard University.

Marianne Mithun. 2000. *Noun and verb in Iroquoian languages: Multicategorisation from multiple criteria*, pages 397–420. De Gruyter Mouton, Berlin, New York.

Marianne Mithun. 2005. *"Routledge Encyclopedia of Linguistics"*, chapter "Mohawk and the Iroquoian languages". New York: Routledge.

Garrett Nicolai, Eleanor Chodroff, Frederic Mailhot, and Çağrı Çöltekin, editors. 2023. *Proceedings of the 20th SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology*. Association for Computational Linguistics, Toronto, Canada.

Hyunji Hayley Park, Katherine J. Zhang, Coleman Haley, Kenneth Steimel, Han Liu, and Lane Schwartz. 2021. Morphology matters: A multilingual language modeling analysis. *Transactions of the Association for Computational Linguistics*, 9:261–276.

Tiago Pimentel, Maria Ryskina, Sabrina J. Mielke, Shijie Wu, Eleanor Chodroff, Brian Leonard, Garrett Nicolai, Yustinus Ghanggo Ate, Salam Khalifa, Nizar Habash, Charbel El-Khaissi, Omer Goldman, Michael Gasser, William Lane, Matt Coler, Arturo Oncevay, Jaime Rafael Montoya Samame, Gema Celeste Silva Villegas, Adam Ek, Jean-Philippe Bernardy, Andrey Shcherbakov, Aziyana Bayyr-ool, Karina Sheifer, Sofya Ganieva, Matvey Plugaryov, Elena Klyachko, Ali Salehi, Andrew Krizhanovsky, Natalia Krizhanovsky, Clara Vania, Sardana Ivanova, Aelita Salchak, Christopher Straughn, Zoey Liu, Jonathan North Washington, Duygu Ataman, Witold Kieraś, Marcin Woliński, Totok Suhardijanto, Niklas Stoehr, Zahroh Nuriah, Shyam Ratan, Francis M. Tyers, Edoardo M. Ponti, Grant Aiton, Richard J. Hatcher, Emily Prud'hommeaux, Ritesh Kumar, Mans Hulden, Botond Barta, Dorina Lakatos, Gábor Szolnok, Judit Ács, Mohit Raj, David Yarowsky, Ryan Cotterell, Ben Ambridge, and Ekaterina Vylomova. 2021. SIGMORPHON 2021 shared task

on morphological reinflection: Generalization across languages. In *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 229–259, Online. Association for Computational Linguistics.

Catherine Price, Keith Lickers, and Karin Michelson. 2011. Native languages. a support document for the teaching of language patterns. oneida, cayuga, and mohawk. The Ontario Curriculum Grades 1 to 12. The Ontario Ministry of Education Resource Guide.

H. Reichenbach. 1947. *Elements of Symbolic Logic*. A Free Press paperback : philosophy. Macmillan Company.

Martin Renard. 2022. Revitalising kanyen'kéha on the grand river: A case study of indigenous language revitalisation and its theoretical implications. *Journal of Undergraduate Linguistics Association of Britain*, 1(2):32–64.

John Sylak-Glassman. 2016. The composition and use of the universal morphological feature schema (unimorph schema).

Clara Vania, Andreas Grivas, and Adam Lopez. 2018. What do character-level models learn about morphology? the case of dependency parsing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2573–2583, Brussels, Belgium. Association for Computational Linguistics.

Shijie Wu, Ryan Cotterell, and Mans Hulden. 2021. Applying the transformer to character-level transduction. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1901–1907, Online. Association for Computational Linguistics.

A Acronyms Used in Linguistic Glosses

Acronym	Explanation
1	first person
2	second person
3	third person
A	agent
BEN	benefective
CIS	cislocative
CONT	continuative
DIST	distributive
DU	dual
DUP	duplicative
FACT	factual
FI	feminine indefinite
FUT	future
FZ	feminine zoic
HAB	habitual
INCH	inchoative
INCL	inclusive
JR	joiner vowel, link
M	masculine
NS	non-singular
OPT	optative
P	patient
PART	partative
PL	plural
PROG	progressive
PUNC	punctual aspect
REP	repetetive
SG	singular
STAT	stative
TRANSL	translocative

Table 4: List of acronyms used in linguistic glosses in the examples.