Three-Step Parsing in Kanien'kéha

Max Blackburn

Department of Linguistics, McGill University

LING 488: Independent Study

Professor Jessica Coon

Abstract

Morphological parsing is the task of transforming a surface linguistic form into its underlying morphological components. Parsing is an essential part of studying morphologically complex languages, where surface words can obscure the underlying linguistic structure. Many such languages are endangered and/or under-resourced, which restricts the usage of the more common data-driven methods in NLP that could automate this task. As a result, most parsing programs are implemented using symbolic finite-state machines. One of the most common architectures for finite-state models of morphology uses a two-step process, which first defines grammatical morpheme sequences, then applies a series of morphophonological rules to derive a surface form (Beemer et al., 2020; Koskenniemi, 1986). I propose a three-step architecture that divides morphophonological and phonological alternations. I claim that this addition constrains the power of the model in a way that mirrors predictions of theoretical linguistics on the structure of morphophonology. As a demonstration of these claims, I implement a three-step parser for verbs in Kanien'kéha, a morphologically complex and highly endangered language. I argue that the results demonstrate that this constrained architecture is still powerful enough to model the language and describe some theoretical findings of the structure of the language. I also further discuss the theoretical and practical questions raised by the choice of model architecture.

1 Introduction

Morphological parsing is the task of extracting morphological information from a surface form in natural language. It is an essential part of analyzing morphologically complex languages: that is, languages with a high number of morphemes per word.[12] It is a routine part of doing linguistic work with such languages. Figure 1 schematizes the task, applied to a word from Kolyma Yukaghir (Maslova, 2003).[3]

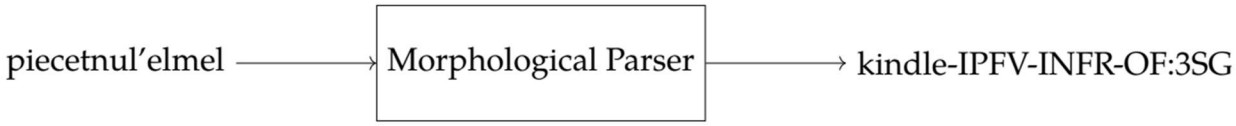

Figure 1: Representation of Morphological Parsing

Many morphologically complex languages are under-resourced. This makes it difficult to use traditional data-intensive approaches for NLP tasks in these languages (Kazantseva et al., 2018). As a result, tools dealing with morphological tasks in these languages are often implemented with a **Finite-State Transducer** (FST), a construct which allows direct, symbolic representation of the language's morphophonology. The FST also allows the usage of generative models as parsers.

The most common architecture for these FSTs divides the modeling process into two components or steps (Beemer et al., 2020; Hulden, 2009; Kazantseva et al., 2018). The first is the lexicon, which is

[1] *Niawenhkó:wa* to Wari McDonald and Akwiratékha' Martin for sharing their knowledge of Kanien'kéha with me and the rest of the McGill Linguistics department. *Niáwen* to Wari McDonald and to Wishe Mittelstaedt for teaching me to speak what I can. Thank you to Akwiratékha' Martin, Anna Kazantseva, Chase Boles, Heather Goad, and the members of the Roti'nikonhrowá:nens reading group for guidance and feedback throughout my research process. And especially thank you to Jessica Coon, for teaching and guidance, and for supervising the independent study which let me pursue this project.

[2] For the purposes of this task, I take a word to an *orthographic* word: that is, a string delimited by punctuation or whitespace. It is within these words that morpheme boundaries are not clearly defined by orthography, making it the domain in which parsing is required. I make no claims regarding the relationship between this word and the phonological or morphological word.

[3] Morpheme breakdowns follow the Leipzig Glossing Conventions. Abbreviations are as follows: AG — agentive, BEN — benefactive, DUP — duplicative, FACT — factual, HAB — habitual, JR — joiner, NEG — negation, NMLZ — nominalizer, NP — noun prefix, NSF — noun suffix, OPT — optative, PUNC — punctual, REP — repetitive, STAT — stative, TRANS — translocative. The pronominal prefixes, described in Section 4.2, distinguish the following features: 1 — first person, 2 — second person, SG — singular, DU — dual, PL – plural, INCL — inclusive, EXCL — exclusive, M — masculine, F — feminine, N — neuter, I — indefinite; agent prefixes are indicated with A, patient prefixes are indicated with P, and the transitive prefixes use > to separate the subject and object respectively.

responsible for (a) defining grammatical sequences of morphemes and (b) replacing morphosyntactic forms with corresponding phonological forms (analogous to the process of Vocabulary Insertion in Distributed Morphology (Halle and Marantz, 1994)). The second is the morphophonology, which transforms the strings of morpheme forms into a surface form, by applying a series of morphophonological rules, illustrated in Figure 2 with the word "wishes".

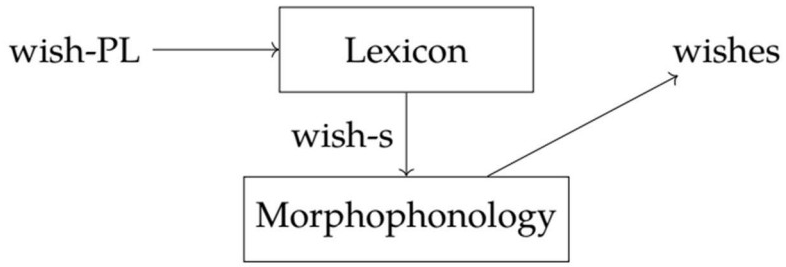

Figure 2: Two-Step Model

While this approach is robust, and many comprehensive finite-state models have been implemented using it (Dunham, 2014; Kazantseva et al., 2018; Lachler et al., 2018; Snoek et al., 2014; Zueva, Kuznetsova, and Tyers, 2020), it makes several implicit assumptions, chief among them, that each morpheme has a default phonological form, which are all inserted simultaneously. This simultaneous insertion complicates the modeling of **phonologically-conditioned allomorphy**, a phenomenon that arises when different allomorphs occur depend ing on phonological context; but, the alternation cannot be explained by phonology alone (i.e., it is suppletive) (Rolle, 2023). For example, in Kanien'kéha (Iroquoian), the neuter agent pronominal prefix has the form *ka-* in most cases, but *w-* before a verb stem beginning with *a*, as demonstrated in (1) and (2).

(1) **ka**hiá:tons
ka-hiaton-s
NA-grab-HAB
'She writes' (Martin 2023:68)

(2) **w**atá:wens
w-atawen-s
NA-swim-HAB
'She swims' (Martin 2023:68)

This alternation is consistently conditioned by the first segment of the verb throughout the language, meaning the trigger is phonological. Additionally, there is no *ka* ∼ *w* alternation in other environments, meaning it cannot arise from phonology.

As an alternative architecture, I propose a three-step model. The first step models the morphology: it defines valid orders of morphemes and other syntactic dependencies. The second step transforms morphemes into their appropriate phono logical forms. The third step represents the phonology: it applies the phonological transformations necessary to derive surface forms. Crucially, the morphology does not make reference to phonology, and the phonology does not make reference to morphology. The second step is entirely responsible for inserting phonological forms: by conditioning these rules, a significant amount of allomorphy can be captured at this step instead of in the phonology. Figure 3 schematizes this with the same word: wishes.

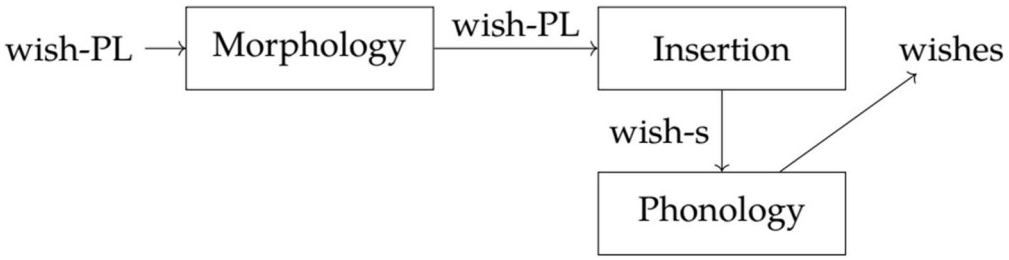

Figure 3: Three-Step Model

The change models a cross-linguistic generalization predicted by certain linguistic models of morphophonology: **phonologically-conditioned allomorphy is almost always conditioned towards the root** (Rolle, 2023; Rolle and Bickmore, 2022).[4] Essentially, if morphosyntactic information is replaced with phonological information from the root outwards, phonological conditioning for spellout will only be available on the side of a morpheme closest to the root (Gouskova and Bobaljik, 2019): Step 2 of Figure 6 shows an example of this operation. Spellout enforces separation of morphosyntactic and phonological information, and constrains suppletive allomorphy.

[4] This pattern is predicted to be universal, though some phenomena have been argued to break it (Rolle, 2023). Regardless of whether it is a near-universal or a true universal, I believe it is still strong enough to inform a choice of model architecture.

As a test for whether these constraints still allow sufficient power for a language model, I implement a parser for basic verbs in Kanien'kéha.[5] I argue that the three-step architecture is still sufficient to represent the language, and that it carries several benefits. In attempting to model the language with these constraints, and building off of previous theoretical and descriptive work, I also present some expanded analyses of Kanien'kéha morphology.

This paper is organized as follows. In Section 2, I describe the problems as sociated with morphological parsing in low-resource languages, and compare the two-step and three-step architectures for finite-state modeling. In Section 3, I give background on Kanien'kéha, the language of interest. In Section 4, I describe specific results of modeling verbs in Kanien'kéha, centered around linguistic findings. In Section 5, I discuss the theoretical and practical generalizations arising from the creation of the parser, and discuss the results of using a three-step architecture. Section 6 concludes.

2 Morphological Parsing

2.1 Problem Description

Morphological parsing is the task of extracting morphological information from surface forms. Since implementation of this task can vary, I define the goal of my parser as being able to produce strings of morphemes, such that each morpheme maps to a separable phonological form.[6] More practically, the goal of this parser is to specifically replicate the manner of parsing performed by linguists, mapping the first line of a standard Inter-Linear Gloss to the third line, as counted in (3).[7]

(3) reader
read-er
read-AG
'reader'

[5] What counts as "basic", or even a "verb" for the purposes of the parser, is defined in Section 3.2.

[6] There are some cases where morphemes can appear fused with others. These are glossed with a period, per Leipzig glossing conventions. These are not to be confused with the morphemes that appear as a fusion of features that are not analyzed as separate synchronically, like the transitive agreement prefixes.

[7] Due to the nature of the finite-state architecture, segmenting the word into phonological strings is not possible without first identifying the morphemes that these strings represent. As such, while extracting these phonological forms of the morphemes (as in the second line of the example) is a useful task, it must follow from the extraction of the morphemes themselves.

2.2 The Finite-State Transducer

The Finite-State Transducer (FST) is a common construct used for morphological parsing or morphological analysis in general. It has the advantage of allowing direct implementation of morphological and phonological rules. These rules can model the analyses created by linguists (Dunham, 2014), meaning that large amounts of data are not required to make a functioning system, as opposed to many of the standard techniques in NLP. As a result, the FST is ideal for modeling under-resourced languages.

FSTs also have the property of reversibility. So, a well-constructed generator FST will also function as a recognizer or parser. Since most linguistic analysis is done from the generative perspective: deriving surface forms from underlying structures, the reversibility property is advantageous. As a result, making a parser only requires implementation of a generative model. In this paper, I will be discussing FSTs and their corresponding linguistic analyses as generative models.

2.3 Two-Step Parsing

The contemporary architecture for finite-state morphological models involves two steps. Both are reflected in the functionalities provided in finite-state toolkits like *foma* (Hulden, 2009), and in the structure of contemporary finite-state models, such as those presented in Kazantseva et al., 2018; Lachler et al., 2018; Snoek et al., 2014.

The first step is variously referred to as the lexicon or the lexical layer, which is responsible for defining grammatical sequences of morphemes (and potentially other subword units, like annotations) (Koskenniemi, 1986).

The second step is the phonology or morphophonology. It is responsible for applying phonological rules. Depending on the language and architecture, it may be responsible for applying morphophonological rules: that is, rules that need to be specified for certain morphological contexts. Certain models, like those in Dun ham, 2014 and Kazantseva et al., 2018, apply all morphophonological rules in this step. Others, like Snoek et al., 2014, divide the process by listing suppletive forms within the lexicon whilst handling more regular morphophonological alternations within the phonology. Figure 2 schematizes this two-step division.

2.4 Components of a Parser

In examining these models, the task of any morphological model can be divided into three components. These components can be viewed as an abstract process, or from two other complementary perspectives. These are their practical implementations in a computational system, and their theoretical equivalents in formal models of linguistics. These are summarized in Table 1.

Each component is described in further detail in Sections 2.4.1-2.4.3.

2.4.1 Morphology

The morphology is responsible for defining valid sequences of morphemes through defining lists of morphemes, and their relationships to each other. Linguistically, this might correspond to a morphological template; computationally, it might be implemented as a large regular expression.[8]

	Morphology	Phonology	Morphology-Phonology Mapping
Role	Defining Grammatical Word Structure	Enforcing Phonological Processes	Mapping Abstract Morphemes to Phonological Equivalents
Implementation	*lexc* Tree, Regular Expression	Rewrite Rules, *TwoLC* constraints	Various
Linguistic Theory	Morphological Template, Generative Grammar	Phonological Rules, Optimality Theoretic Constraints	Distributed Morphology Vocabulary Insertion Rules

Table 1: Components of a Morphological Model

[8] The suitability of morphological templates for this purpose is addressed in 5.3.

2.4.2 Phonology

The phonology models phonological and orthographic rules that are applied as a result of the morphemes being concatenated. Linguistically, this might correspond to a series of ordered generative phonological rules. Computationally, this might be implemented using a cascade of rewrite rules.[9]

2.4.3 Morphology-Phonology Mapping

Since the morphology is defined in terms of abstract morphemes, and the phonology is defined in terms of phonological strings, the model must have a way of mapping between morphemes and their phonological forms. In the framework of Distributed Morphology (Halle and Marantz, 1994), this is analogous to the process of Vocabulary Insertion, or the broader concept of spellout. As such, I adopt the term **spellout** to refer to this process. Like phonology, spellout can be implemented using rewrite rules.

2.5 Spellout in Two-Step Parsing

Spellout in finite-state models is almost always implicitly handled by the lexicon. Figure 4 shows a sample of the *lexc* file presented in Snoek et al., 2014. The line < +Obv:a > ; indicates that the +Obv tag should be spelled out as a.

```
LEXICON ANSTEMLIST
apiscacihkos ANDECL ;
apisimˆosos ANDECL ;
LEXICON ANDECL
< +N:0 +AN:0 +Sg:0 @U.noun.abs@ # > ;
< +N:0 +AN:0 @U.noun.abs@ OBVIATIVE > ;
LEXICON OBVIATIVE
< +Obv:a # > ;
```

Figure 4: Sample of the Plains Cree Lexicon from Snoek et al., 2014

I argue that this model of spellout complicates modeling of **phonologically-conditioned allomorphy** (PCA). Per Rolle, 2023, I take PCA to be allomorphy which is suppletive (i.e., it cannot be attributed to a broader phonological process) and conditioned on the presence of a phonological property. PCA is complicated by the fact that it needs to make reference to both morphological information (the

[9] This can also be implemented using parallel constraints on surface forms (Koskenniemi, 1986). In linguistics, this is analogous to the constraint-based phonology of Optimality Theory (Violin Wigent, 2006).

relevant morpheme) and phonological information (the relevant trigger for the allomorphy). As a result, it must be conditioned at one of two parts of the morphological derivation: either (1) during the process of spellout or (2) after the process of spellout, with morphological information still visible in some fashion. However, the two-step model also makes the following assumptions:

- **Every morpheme has a default allomorph:** by listing pairs of morphemes and allomorphs, these allomorphs are assumed to be default forms.
- **Morphemes are all spelled out simultaneously:** by associating default allomorphs with morphemes in the same structure that morpheme sequences are being defined.

As a result of these assumptions, there is no control during the process of spellout, since all morphemes are spelled out in predetermined forms simultaneously. Option (2) is the only one available for modeling PCA in a two-step model. In other words, the two-step morphological model **requires** morphological information to still be available after spellout is complete. Figure 5 demonstrates this with a derivation of the Kanien'kéha word *watá:wens*. The initial *a* of the verb stem *atawen* conditions the allomorph *w-* for the NA prefix. However, the prefix is instead spelled out with its default form, *ka-*. It must then be readjusted with a specialized rule which requires the presence of morphological annotations:[10]

[10] The phonology is simplified as it is not relevant to the process of spellout. I make no claims about the general structure that is required for phonology, such as overt morpheme boundaries, morphophonological domains, or cyclic rules.

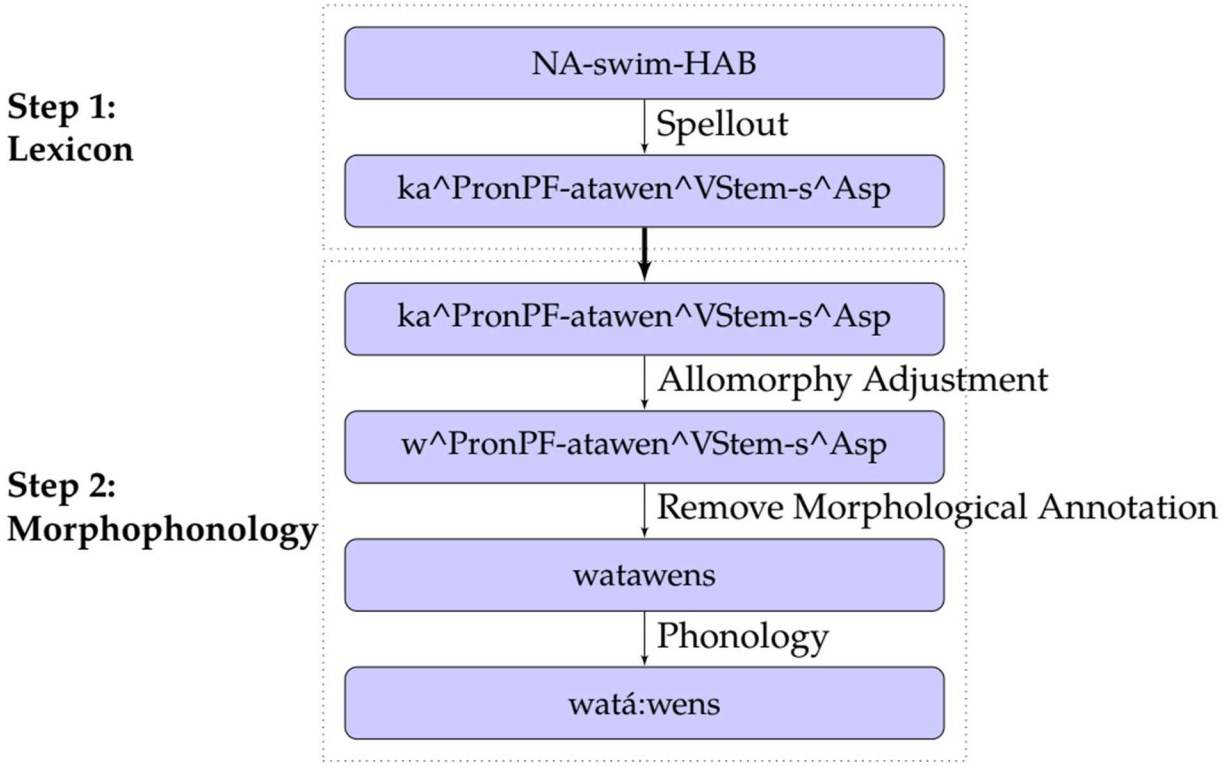

Figure 5: Two-Step Modeling of Phonologically-Conditioned Allomorphy

2.6 Three-Step Parsing

As an alternative, I propose a three-step architecture for morphological models, grounded in the following principles in theoretical linguistics:

1. **Late Insertion**: spellout of morphemes only occurs **after** the entire morphological structure has been built, a core component of Distributed Morphology (Halle and Marantz, 1994).
2. **Cyclic Insertion:** spellout of morphemes occurs in a sequential manner. Per the principles of Distributed Morphology, I assume this begins in the root and proceeds outwards one morpheme at a time (Gouskova and Bobaljik, 2019).

To do this, the three-step architecture aims to separate morphophonological pro cesses, especially phonologically-conditioned allomorphy, from the other components of the model entirely. This is done by

removing all spellout rules from the lexicon, all readjustment rules from the phonology, and defining spellout rules within an entirely new component. The result is a three-step process: the morphology/lexicon, which feeds the spellout, and in turn feeds the phonology. The process is shown in Figure 6, where a three-step model derives the same word, *watá:wens*, that was used in the example in Figure 5:

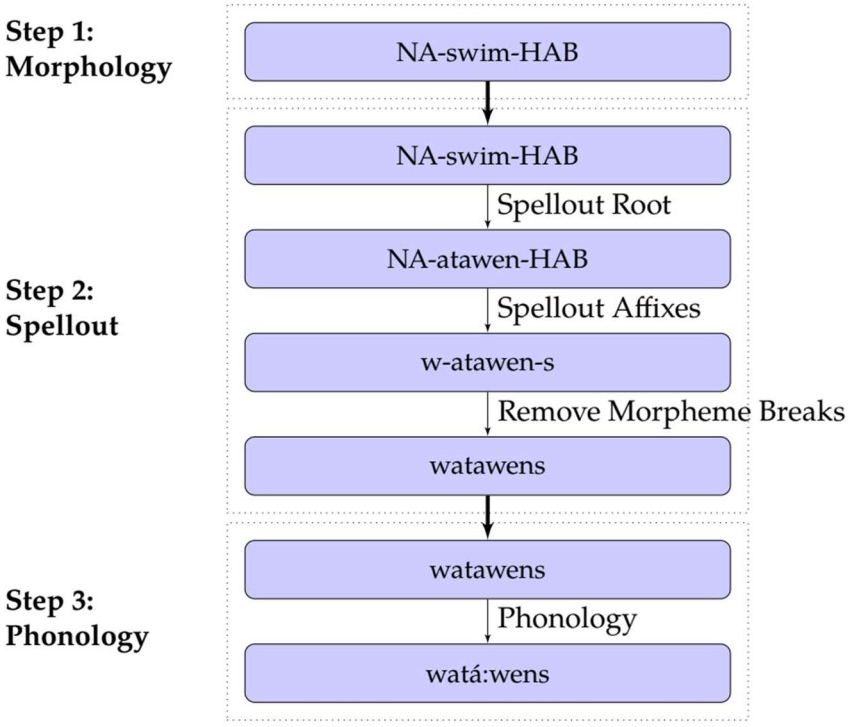

Figure 6: Three-Step Modeling of Phonologically-Conditioned Allomorphy

As seen in this derivation, phonological material is only available towards the root, as it gets gradually inserted. As a result, the correct allomorph of the neuter agent, *w-*, is inserted immediately. The principal benefit of this approach is that it removes redundant suppletive readjustment rules, simplifying the structure of the model.[11]

2.7 Immediate Theoretical Considerations: Phonologically-Conditioned Allomorphy

[11] It is likely that minimization of the FST would result in no practical increases in runtime efficiency. The benefits I argue for here primarily concern program architecture.

An immediate implication of this architecture is that phonologically conditioned allomorphy that is conditioned outwards (away from the root) is absent. This generalization is well supported by the near total lack of such phenomena cross-linguistically, with models like Distributed Morphology predicting them to be totally absent (Rolle, 2023). As a result, I argue that this loss of expressive capability within the model architecture is well-justified, and thus a desirable simplification.

Further theoretical and practical considerations of the three-step architecture are discussed in Section 5.

3 Language Background: Kanien'kéha

3.1 Ethnography

Kanien'kéha (also known as Mohawk) is a language belonging to the Lake Iroquoian branch of the Iroquoian language family (Michelson, 1988). It is spoken by the Kanien'kehá:ka, who today live in territories and communities throughout Ontario, Quebec, and New York State. Kanien'kéha is classified as severely endangered on the UNESCO Intergenerational Transmission Metric, and Moribund or Nearly Extinct on the Graded Intergenerational Disruption Scale (GIDS). This reflects a severe language shift that took place in the 20th century, mostly as a result of government programs like residential schools and forced adoption. Today, like with many other indigenous communities, there is a strong language revitalization movement (DeCaire, 2023).

3.2 Language Description

Kanien'kéha is described as being a polysynthetic language: defined as a language with a high morpheme to word ratio (DeCaire, 2023). Kanien'kéha has three lexical categories: nouns, verbs and particles. Particles are uninflected and usually relatively short. Nouns consist minimally of a noun stem, with a noun prefix and a noun suffix. Verbs contain minimally a verb stem, pronominal agree ment prefix, and aspect suffix (Michelson, 2020). Examples of a noun and verb are shown in (4) and (5) respectively.

(4) ò:niare'
o-hniar-'
NP-snake-NSF
'snake' (Michelson et al. 2024:135)

(5) rón:nis
hr-onni-s
MSGA-make-HAB
‘He makes it.‘ (Martin 2023:68)

Since particles are, by definition, uninflected, they do not require parsing. Nouns have some morphological complexity, but are generally transparent. Verbs present a much broader range of morphosyntactic structures and morphophonological phenomena, as seen in (6).

(6) Wahakeriseranénhsko’
wa-**hak**-**ris**-er-**a**-nenkshw-’
FACT-MSG>1SG-sock-NMLZ-JR-steal-PUNC
‘He stole the socks from me.‘ (McDonald 2024:2952)

This word showcases many characteristic traits of Iroquoian morphology, such as the use of a transitive pronominal prefix *hak* that represents agreement with both the subject and indirect object, noun incorporation of the direct object *ris*, and the insertion of a joiner vowel and an e epenthetic vowel to resolve illegal consonant clusters. Such phenomena are widespread, and cannot be ignored in a description or analysis of the language, be it pedagogical, theoretical, or descriptive. As a result, verbs are an ideal domain to test the full capabilities of automatic parsing.

It should be noted that these categories are defined structurally, not semantically. Words that are structured like verbs can be used referentially, as demonstrated in (7).

(7) Rana’tarón:nis wà:rewe’.
ra-na’tar-onni-s wa-hr-ew-’
MSGA-break-make-HAB FACT-MSGA-arrive-PUNC
‘The baker arrived (lit. He who makes bread arrived)’ (McDonald 2024:850)

For the purposes of defining the target morphological domain in which parsing should be successful, a "verb" is a word which fulfills the structural definition, not the semantic definition. Figure 7 defines the portions of the verb which the current version of the parser recognizes. Elements in parentheses are optional, while the square brackets delimit the domain of the verb stem.

(DUPLICATIVE)-(MODAL)-PRONOMINAL PREFIX-[(**REFLEXIVE**)-**VERB ROOT**]-ASPECT

Figure 7: Target Section of the Kanien'kéha Verb

Some previous work has been done in computational modeling of Kanien'kéha morphology. Alicia Assini created a finite-state parser for nouns in Kanien'kéha. This program was created in *foma* and can parse a subset of nouns in Kanien'kéha (Assini, 2013), similar to my project, except my goal is a parser for verbs.

A team with members from the Onkwawénna Kentyóhkwa Kanyen'kéha im mersion school and the Canadian National Research Council's Indigenous Language Technology lab created Kawennón:nis, a verb conjugator for Kanien'kéha. Designed as an aid for students learning the language, it is capable of conjugating hundreds of verb roots across a variety of combinations of person, aspect, and other important features. At its core is a finite-state transducer, which models the morphological and phonological processes which transform underlying features into surface forms (Kazantseva et al., 2018).

Like all finite-state transducers, this FST can be run in reverse, resulting in a verb parser; however, since it was designed as a generator which works with English translations, doing so causes some issues. One is that there can be semantic ambiguity in the results. For example, parsing watá:wens returns two parses, one marked habitual and one marked state. These presumably correspond to the multiple meanings available in the habitual aspect: however, these both correspond to the same morpheme structure. These could in theory be adapted or discarded in order to make a machine that parses only syntactic information. The more important quality of the Kawennón:nis FST is that it was constructed as a two-layer model. As such, creating a three-layer parser for Kanien'kéha is not a redundant task, given the existence of machine with similar capabilities. Rather, it can serve as a metric for comparing the two architectures.

4 Results

As a symbolic model of the language, practical results of choices in implementing the parser are inextricably linked to theoretical findings of linguistic analysis. Key findings regarding the behavior of specific morphemes are described below.

4.1 Theoretical Assumptions

4.1.1 Morphology

As a result of the finite-state machine constraints, I assume a templatic model of morphology, with a limited number of long-distance dependencies.

4.1.2 Spellout

I assume spellout takes place cyclically, with phonological and morphosyntactic conditioning available to Vocabulary Insertion in line with the constraints introduced in Section 2.

4.1.3 Phonology

I model the phonology as a system of rules with potentially intrinsic ordering, motivated primarily by two factors:

1. Most descriptive and analytic literature on Iroquoian phonology uses a rule based framework. This notably includes Michelson, 1988, upon whose analysis my implementation relies heavily.
2. A rule-based phonology allows mechanisms implemented for Vocabulary Insertion to be reused.[12]

I assume the operation of certain rules to be necessarily constrained by morphophonological domains. Due to the practical constraints of implementing phonological rules in environments with morpheme boundaries, I avoid making reference to morpheme boundaries within rule formulations. However, I do not outright reject their use in this context.

4.2 Pronominal Prefixes

[12] An example is the spreadsheet to source code compiler described in Section 5.2.

As one of the three mandatory parts of any verb in Kanien'kéha, a comprehensive model of pronominal prefixes is required to model the structure of verbs. They can mark agreement in various combinations of person, number, gender, clusivity, and thematic role. These thematic roles can represent an agent (active participant in the verb), patient (less active participant in the verb), or a transitive relation (both active and less active participants). (8), (9), and (10) demonstrate these respectively.

(8) khiá:tons
k-hiaton-s
1SGA-buy-HAB
'I buy (it)' (Martin 2023:68)

(9) wakenoròn:se'
wak-noron's-e'
1SGP-exhausted-HAB
'I am exhausted' (Martin 2023:78)

(10) konhró:ris
kon-hrori-s
1SG>2SG-tell-HAB
'I tell you' (Martin 2023:72)

These prefixes are also subject to pervasive phonologically-conditioned allomorphy, where the first segment of the verb stem can condition suppletive forms.[13] As demonstrated in previous sections, the neuter agent pronominal prefix can occur variably as *ka-* before C-stems or *w-* before A-stems. Some allomorphs can be analyzed as a result of synchronic phonology. In trying to explore the exact mechanics of phonologically-conditioned allomorphy, I have attempted to synchronically analyze each set of allomorphs such that they are small as possible, reducing alternations to provably synchronic phonological processes.

Descriptions of stem classes vary in linguistic and pedagogical literature, counting as many as 9 (McDonald, 2023a). Most settle on 5, however: **C-**, **A-**, **I-**, **E-**, and **O-**stems (DeCaire, 2023; Martin, 2023; Michelson, 1988). Allomorphs are listed exhaustively for the sake of symmetry, and often for ease

[13] This segment can be a part of an incorporated noun, the verb root, or a reflexive or semi reflexive.

of teaching. However, even in these analyses there are many identical forms. With a fine-grained analysis, and additional well-motivated phonological rules, most can be collapsed into one or two sets of allomorphs. These phonological rules include:[14]

1. **Hiatus resolution** rules based on Hopkins, 1987 such as /ai/ → [ɶ̃] and a variety of deletion rules.
2. **Hiatus resolution** rules based on Michelson, 1988 such as /w/ → [j] */ V _ V*.
3. **E-epenthesis** rules from Michelson, 1988.
4. **Coda glide deletion:** /w/, /j/ → *∅ / V _ C*. This is motivated by observations of corpus data: no glides were observed in this environment.

For example, the **1incl>FI** is usually listed with three allomorphs: *iethi*, *iethii*, and *ieth*. These can be unified into one allomorph with form /jethij/, which undergoes glide deletion before C-stems, or experiences the changes in (11) before I-stems:

(11) /jethij-i/ … ——————————> /jethi-i/ ——————————————————> /jethi/

Deletion of /j/ before /i/ Deletion of identical sequences of vowels

A full list of allomorphs from the analysis can be found in Appendix A. The prefixes with three or more forms are the FI.A, MPLA, FPLA, 2SG>1SG, 2SG>1DU, 2SG>1PL, and MSG>1DU.

4.2.1 Second Person Transitives

Certain second person transitive pronouns are subject to another layer of allomorphy. Specifically, those with a *ta-* element have this element change to *hs-* when material is introduced before the prefix (Martin, 2023). Examples (12) and (13) characterize this alternation.

[14] These rules are formulated using IPA notation. In this section, I use IPA in contexts where orthography would be ambiguous, such as underlying forms where *enV* could represent a vowel nasal-vowel sequence or a sequence of a nasal vowel and a vowel in hiatus. The IPA and orthog raphy are identical, except in the following cases: <*i*> before a vowel is the palatal glide /j/, the nasal vowels <*en*> and <*on*> are /ɶ̃/ and /ũ/, and the glottal stop <*'*> is /ʔ/. IPA is indicated with slashes // or brackets [].

(12) takhró:ris
tak-hrori-s
2SG>1SG-tell-HAB
'You tell me' (Martin 2023:66)

(13) iah tehskehró:ris
iah te-hsk-hrori-s
NEG NEG-2SG>1SG-tell-HAB
'You don't tell me' (Martin 2023:66)

These are also sometimes conditioned by the stem class. This can lead to two dimensional variation, as seen in Table 2 with the 2SG>1SG prefix.

	Preceding Material?	
Stem Class	No	Yes
{a}	takw-	hskw-
Elsewhere	tak-	hsk-

Table 2: 2nd Person to 1st Person Allomorphs

The extra dimension is of note, since it constitutes an example of **outwardly conditioned allomorphy**. In accordance with the generalization that phonologically conditioned allomorphy is always inward-conditioned, we should expect the trigger of this alternation to be the presence of syntactic material, and not phonological material. This is borne out by imperative forms like (14): the pronominal prefixes optionally undergo the same alternation, despite the absence of any overt preceding material.[15][16]

(14) skhní:non's
hsk-hninon-'s
2SG>1SG-buy-BEN

[15] These forms are idiolectal in Kanien'kéha, but are common in Oneida (Lounsbury, 1953).

[16] Preliminary evidence indicates that there is a difference in meaning between the forms in *tak* and *hsk*-, making this not a purely optional alternation (Chase A. Boles, p.c.).

‘Buy (it) for me‘ (McDonald 2023)

(14) suggests the presence of some sort of imperative prepronominal prefix, which is always spelled out as null. Indeed, this was the simplest way to encode this into the parser architecture. However, for the purposes of dissecting allomorphy, the important takeaway is that this allomorphy cannot be conditioned phonologically. Since it is sensitive to a distinction between the absence of a preceding element, and a phonologically null preceding element, a phonological trigger would require a distinction between two types of “null” phonology; an unappealing prospect. Therefore, the trigger must be syntactic, in line with the generalizations on allomorphy described.

4.2.2 Interim Conclusions

In summary, in applying a broad enough set of well-motivated phonological pro cesses, the allomorphic variation in the pronominal prefixes can be greatly reduced. These reduced sets of allomorphs correspond to natural classes along with an “elsewhere” case. Rare cases of outwardly-triggered allomorphy are based on syntactic elements, in line with cross-linguistic trends in allomorphy.

Given the complexity of the system of pronominal prefixes and the attention given to them in Iroquoian linguistics, this analysis is likely incomplete in some aspects. The discussion of whether the prefixes can be synchronically divided into agent and patient elements is present in many works (c.f. Hopkins, 1987). I do not attempt this question because it complicates modeling of phonological processes, especially hiatus resolution rules. Additionally, transitive prefixes are not always easily separable into agent and patient components, suggesting morpheme fusion. An analysis breaking down these components would need to address synchronic alternations between fusion and non-fusion. This would also require a theory on whether the fusion is a result of morphosyntactic factors (such as interactions between Phi-features) or morphophonological factors (interactions between morphemes during Vocabulary Insertion). I present an analysis of once instance of morpheme fusion between pre-pronominal prefixes in Section 4.3.3. Examination of fusion phenomena within this less complex domain might provide answers about the patterns of fusion that should be expected within pronominal prefixes.

4.3 Modal Prefixes

Modal prefixes, which occur before the pronominal prefixes, are also subject to a significant degree of allomorphy. Their analysis in this section relies strongly upon the descriptions of this variation in Martin, 2023.

4.3.1 Factual

The factual has the form *we-* before *s* and *t*, shown in (15) and (16):

(15) Nahó:ten **we**senihní:non'?
nahoten we-sni-hninon-'
what FACT-2DUA-buy-PUNC
'What did you buy?' (McDonald, 2024)

(16) Nahó:ten **we**tewahní:non'?
nahoten we-twa'-hninon-'
what FACT-2INCLA-buy-PUNC
'What did we buy?' (McDonald, 2024)

As in (17) and (18), it has the form *wa'-*, with a glottal stop, before *i* (/j/) and *k*:

(17) wa'kón:ni'
wa'-k-onni-'
FACT-1SGA-make-PUNC
'I did make it' (Martin 2023:95)

(18) wa'ehiá:ton'
wa'-ie-hiaton-'
FACT-FI.A-write-PUNC
'She did write' (Martin 2023:96)

Elsewhere, the factual surfaces as *wa-*. However, under the assumption that allomorphy is conditioned by natural classes, positing *wa-* as the default morpheme would require stipulating {/k/,/j/} be a natural class, which is difficult to argue. As such, *wa'-* must be the default allomorph, with *wa-* surfacing before *w*. The conditioning environments are summarized in Table 3.

	Preceding Material?	
Prefix Class	No	Yes
{s,t}	we-	?
{w}	wa-	
Elsewhere	wa'-	

Table 3: Factual Allomorphs

Since modals can only occur before pronominal prefixes, which are limited in their initial segments, the "elsewhere" case is limited to the segment set {/k/,/j/,/r/}. The first two are accounted for. *r* surfaces as *h* with preceding material, ensuring that the factual will only be prefixed to an *h*. To explain the surface forms in *wah-*, all that remains is to posit a laryngeal reduction rule of /P/ $\longrightarrow \emptyset$ / /h/ (Michelson, 1988).

4.3.2 Optative and Future

The future has one form, *en-* (Bonvillain, 1973; Martin, 2023). The optative has two forms: the default is *aa-*.[17] Another form surfaces before {s,t}, which can be variable, demonstrated by (19) and (20):

(19) **ae**sakonhré:konke'
ae-sa-konhrek-on-k-'
OPT-2SGP-hit-STAT-CONT-PUNC
'You should have hit it' (Michelson 1988:37)

(20) **aie**tsatekhón:ni'
aie-ts-at-kh-onni-'
OPT-2DUA-SRFL-food-make-PUNC
'Yous d. ought to have a good meal' (Martin 2023:100)

[17] See Hopkins, 1987 for motivating the underlyingly long form.

The variation between *ae-* and *aie-* can be derivationally linked by a process which epenthesizes /j/ in the vowel sequence: ∅ −→ /j/ / a e. With these forms reduced, a striking parallel with the factual emerges, demonstrated by the allomorph table in Table 4.

	Modal	
Prefix Class	Optative	Factual
{/s,t/}	/ae/	/we/
Elsewhere	/aa/	/wa(ʔ)/

Table 4: Parallel Optative and Factual Allomorphs

This suggests a phonological process is responsible for these alternations, though whether it is synchronic remains to be seen.

4.3.3 Duplicative + Factual

Due to its occurrence before a large number of verbs, the duplicative was the sole non-modal prepronominal prefix I investigated extensively in this analysis. (21) and (22) show its form before a consonant and vowel respectively:

(21) **te**satonhontsó:ni
te-sa-atonhontsoni
DUP-2SGP-want/need.STAT?
'You want/need' (Martin 2023:84)

(22) **ta**hsenónniahkwe'
t-aa-hs-nonniahkw-'
DUP-OPT-2SGA-dance-PUNC
'You ought to dance' (Martin 2023:86)

In (21), <ts> is a valid consonant cluster in initial position, so the <e> must not be epenthetic. However, if the <e> were underlyingly present in (22), the hiatus resolution rule adopted from Hopkins, 1987 would predict a surface form of **tehs-*. This motivates an analysis where the duplicative has the

allomorph *te-* before consonants and *t-* before vowels. However, as seen in (23) and (24), the introduction of the factual causes several puzzling effects.

(23) wa'tkennónniahkwe'
wa'-t-k-nonniahkw-'
FACT?-DUP?-1SGA-dance-PUNC
'I danced' (Martin 2023:84)

(24) wa'tisatonhóntsohse'
wa'-ti-sa-atonhontsohs-'
FACT?-DUP?-2SGP-want?-PUNC
'You did want it' (Martin 2023:85)

(25) wa'tisatonhóntsohse'
wa'-ti-sa-atonhontsohs-'
FACT?-DUP?-2SGP-want?-PUNC
'You did want it' (Martin 2023:85)

First, the positions of the modal and the duplicative have swapped. Whereas the duplicative preceded the optative (and the future (Martin, 2023)), it follows the factual. Additionally, the factual does not have the expected form: before a /t/, we would expect it to be realized as /we/-. Finally, the duplicative also does not have the expected -/e/- before consonants, and has the form /ti/- before /s/.

I argue that this puzzle can be resolved by analyzing the *wa't(i)-* forms as a fusional form: it is not actually separable into a factual and a duplicative. The allomorphs would then be those listed in Table 5.

	Morpheme(s)	
Environment	Duplicative	Duplicative-Factual
{a,e,i,o,ʌ̃,ũ}	/t/	
{/s,t/}		/waʔti/
Elsewhere	/te/	/waʔt/

Table 5: Duplicative and Duplicative+Factual Allomorphs

Under this analysis, the modal and duplicative do not swap positions: they combine in certain cases. This allows for a simplification of syntactic analysis: for example, some morpheme templates, like those in Bonvillain, 1973, have an extra slot containing just the factual, seemingly only to resolve this one case.

This analysis also simplifies phonological analysis. By positing an underlying /e/ only in the non-fused form, the distribution of surface [e] becomes explainable. Michelson, 1988 mentions that it is not clear whether the duplicative contains an epenthetic [e] or not, since the form *te-* occurs in cases where epenthesis is not predicted, but that in other cases it does not surface at all. In the examples below, (26) contains an epenthetic vowel as a result of the *-'tkt* cluster, while (27) does not, since *-'tk-* is a valid cluster.

(26) wa'tekté:ni'
wa't-k-teni-'
FACT.DUP-1SGA-change-PUNC
'I changed (it)' (Michelson 1988:135)

(27) wa'tkené :ra'ke'
wa't-k-nera'k-'
FACT.DUP-1SGA-mistake-PUNC
'I mistook (it)' (Michelson 1988:136)

In ignoring the duplicative forms that occur without the factual, the surface in stances of [e] in the context of the factual can be derived using e-epenthesis rules.

5 Discussion

Having presented the architecture of the three-step model, as well as the linguistic analyses that arise from its constraints, I now discuss its other theoretical and practical consequences.

5.1 Implementation

I implemented my model in *foma*, a finite-state toolkit (Hulden, 2009). *Foma* provides the full functionalities required to construct finite-state machines from regular expressions and rewrite rules. It natively supports the usage of .lexc files, which are a common format for defining the lexicon step in a two-step program.

Implementing a three-step parser within this framework is an approach that only requires two components:

1. The lexicon must not spellout morphemes. This allows the spellout com ponent full control over allomorphy without resorting to later readjustment rules. It may be dispensed with if the relevant morphemes do not participate in or trigger allomorphy.
2. If the usage of a single .lexc and a single .foma file is required, the spell out can be included in the .foma file, composed before the phonology. It is also helpful to define filters that ensure that the process of spellout is complete in the generative and recognition directions, since the separation of morphology and spellout does not guarantee that morphemes are converted to phonology. A sample pattern for a filter for this output would be [C | V | "-"]*, assuming that consonants and vowels are defined, and that morpheme breaks should be available to the phonology.

5.2 Allomorphy Constraints

Since the three-step framework avoids readjustment rules, allomorphy conditioning is much more limited in its scope. Outwardly-conditioned allomorphy can only be grammatical, and inwardly-conditioned allomorphy is potentially only phonological. Regardless of the framework used, this allows for these conditioning variables to be formalized. For example, sets of allomorphs, inward conditioning, and outward conditioning could be represented by the following 3-tuple, where P is the segmental inventory of the language and G is the set of grammatical features conditioning allomorphy:

(28) $\{(a \in P^*, i \in (G \cup C \text{ for } C \subseteq P), o \in \{T, F\}) : C \text{ is a natural class}\}$

This particular formulation predicts that inward-conditioned allomorphy can be sensitive to a grammatical feature or a natural class of segments, while outwardly conditioned allomorphy is

conditioned on the binary presence of a syntactic element (as seen with the prefixes in Table 2).[18] I was able to use this consistency to store lists of allomorphs in a spreadsheet, which could then be automatically compiled to *foma* using a python script. This spreadsheet format allowed for easier data storage and visualization, and avoided errors in the development process that might be triggered by working directly with the *foma* source code. Spreadsheets are one of the most widely used tools for storing and organizing linguistic data: as such, this also improves cross-compatibility between sources of data, and avoids the destruction of legible linguistic data when encoding it into the source code of the program (Littell et al., 2024).

5.3 Morphosyntactic and Semantic Dependencies

Verbs in Kanien'kéha exhibit many features that reflect morphosyntactic and semantic dependencies, often over long distances: for example, modal prefixes only occur with a punctual suffix. I attempted to model this phenomenon and others in the morphology layer of my parser, and was most successful through defining several alternative templates for specific paradigms. This reflects a key limitation in finite-state modeling of morphosyntax: finite-state machines can only model regular languages (Koskenniemi, 1986), whereas natural language morphosyntax is minimally context-free and not regular (Chomsky, 1956). The mirror of this in linguistics is the general finding that morphological templates are insufficient theoretical tools for modeling and prediction (Crippen, 2019).

5.4 Separating Morphology

As a consequence of separating spellout from the morphological component, the latter is not strictly required for generation. Since the spellout and phonology are entirely capable of transforming a morpheme string to a surface form, the morphology can be replaced by another generative component. As mentioned in Section 5.3, this could be desirable, as a more adequately powerful model of morphosyntax could be employed instead of the limited template. This would reflect a modular theory of linguistics, where the output of the syntax feeds the (morpho-)phonology.

However, this optionality of the morphological component leads to an interesting result. In the parsing or recognition direction, potential underlying forms need to be pruned out by some morphological component. Since the pruning mechanism must be able to be concatenated and then minimized with the other components to avoid overgeneration, it must be regular and not context-free. This contradicts the

[18] As demonstrated by the list of segment sets conditioning allomorphy in Appendix A and the corresponding natural classes in Appendix B, phonologically conditioning segments form natural classes, as expected by the principle that phonology is sensitive to features and not segments.

necessity of a minimally context-free model of morphosyntax. Thus, one of the following assumptions must be made within a morphological model in order to avoid overgeneration of underlying forms at any point in the derivation:

1. There is a regular grammar, like a morphological template, that is active only when recognizing linguistic material and not generating it.
2. There is a regular grammar, like a morphological template, that is active within the generative and recognition process, which redundantly filters the output of a more powerful generative component.
3. Phonological rules are heavily constrained such that they do not create an exponentially larger number of underlying forms.

Option 3 is potentially the most desirable, since it results in a grammar that is neither redundant nor asymmetric, nor does it appeal to inadequate mechanisms like templates. However, it also requires the most care when implemented practically, since the creation of morphological models is done from the generative direction, and it is quite easy to overlook the potential overgeneration that is available when reversing any given transformation. The addition of a redundant or asymmetric component is therefore a more secure option to avoid overgeneration.

6 Conclusion

I have claimed that the introduction of an intermediate layer in an FST would result in a necessary reduction in expressive power. I have shown that the three-layer architecture is still powerful enough to model a variety of linguistic phenomena, at least in this experiment. The problems it experiences, such as challenges with long-distance dependencies, are a result of the finite-state framework itself. This reduction in expressive power also creates some practical benefits for development of these models, as well as raising theoretical questions of their development being automated. Probing the architecture of these models also raises interesting questions about the components which are necessary both in theoretical models of morphology and their practical implementations.

References

Assini, Alicia Alexandra. 2013. Natural language processing and the Mohawk language: creating a finite state morphological parser of Mohawk formal nouns. Limerick, IE: University of Limerick MA thesis.

Beemer, Sarah et al. 2020. Linguist vs. Machine: Rapid Development of Finite-State Morphological Grammars. *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*. Ed. by Garrett Nicolai, Kyle Gorman, and Ryan Cotterell. Online: Association for Computational Linguistics (ACL). 162–170.

Bonvillain, Nancy. 1973. *A grammar of Akwesasne Mohawk*. University of Ottawa Press.

Chomsky, Noam. 1956. Three models for the description of language. *IRE Transactions on Information Theory* 2(3). 113–124.

Crippen, James A. 2019. The syntax in Tlingit verbs. Vancouver, BC: UBC PhD thesis.

DeCaire, Ryan. 2023. The Role of Adult Immersion in Kanien'keha Revitalization. Hilo, Hawaii: UH Hilo PhD thesis.

Dunham, Joel Robert William. 2014. The online linguistic database : software for linguistic field work. Vancouver, BC: UBC PhD thesis.

Gouskova, Maria and Jonathan David Bobaljik. 2019. Allomorphy and Vocabulary Insertion.

Halle, M and Alec Marantz. 1994. Some key features of distributed morphology. English (US). *MITWPL 21*. Ed. by A Carnie, H Harley, and T Bures. 275–288.

Hopkins, Alice W. 1987. Vowel Dominance in Mohawk. *International Journal of American Linguistics* 53(4). 445–59.

Hulden, Mans. 2009. Foma: a Finite-State Compiler and Library. *Proceedings of the Demonstrations Session at EACL 2009*. Ed. by Jörn Kreutel. Athens, Greece: Association for Computational Linguistics. 29–32.

Kazantseva, Anna, Owennatekha Brian Maracle, Ronkwe'tiyóhstha Josiah Maracle, and Aidan Pine. 2018. Kawennón:nis: the Wordmaker for Kanyen'kéha. *Proceedings of the Workshop on Computational Modeling of Polysynthetic Languages*. Ed. by Judith L. Klavans. Santa Fe, New Mexico, USA: Association for Computational Linguistics. 53–64.

Koskenniemi, Kimmo. 1986. Compilation of automata from morphological two-level rules. *Proceedings of the 5th Nordic Conference of Computational Linguistics (NODALIDA 1985)*. Ed. by Fred Karlsson. Helsinki, Finland: Department of General Linguistics, University of Helsinki, Finland. 143–149.

Lachler, Jordan, Lene Antonsen, Trond Trosterud, Sjur Moshagen, and Antti Arppe. 2018. Modeling Northern Haida Verb Morphology. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Ed. by Nicoletta Calzolari et al. Miyazaki, Japan: European Language Resources Association (ELRA).

Littell, Patrick, Darlene A. Stewart, Fineen Davis, Aidan Pine, and Roland Kuhn. 2024. Gramble: A Tabular Programming Language for Collaborative Linguistic Modeling. *International Conference on Language Resources and Evaluation*.

Lounsbury, Floyd Glenn. 1953. *Oneida verb morphology / Floyd G. Lounsbury.* eng. Yale University publications in anthropology no. 48. New Haven: Published for the Dept. of Anthropology, Yale University, by the Yale University Press.

Martin, Akwiratékha'. 2023. *Tekawennahsonterónnion Kanien'kéha Morphology*. Edition 2. Kanien'kehá:ka Onkwawén:na' Raotitióhkwa' Language and Cultural Center.

Maslova, Elena. 2003. *A Grammar of Kolyma Yukaghir*. Berlin, New York: De Gruyter Mouton.

McDonald, Mary Onwá:ri Tekahawáhkwen. 2023a. McGill Slides 5 - Verb Forms. Slides created for teaching Kanien'kéha.

McDonald, Mary Onwá:ri Tekahawáhkwen. 2023b. McGill Slides Objective. Slides created for teaching Kanien'kéha.

McDonald, Mary Onwá:ri Tekahawáhkwen. 2024. Interviews. Kanien'kéha language consultation work with McGill Linguistics.

Michelson, Gunther, Michelson, Karin, & Canadian Deer, Glenda. 2024. *A Dictionary of Kanien'kéha (Mohawk) with Connections to the Past*. University of Toronto Press.

Michelson, Karin. 1988. *Lake Iroquoian accent*. Kluwer.

Michelson, Karin. 2020. Word Classes in Iroquoian Languages. *The Oxford Handbook of Word Classes*. 651–668.

Rolle, Nicholas. 2023. Inward and Outward Allomorph Selection. *The Wiley Blackwell Companion to Morphology*. 1–30.

Rolle, Nicholas and Lee Bickmore. 2022. Outward-sensitive phonologically-conditioned suppletive allomorphy vs. first-last tone harmony in Cilungu. *Morphology* 32. 197–247.

Snoek, Conor, Dorothy Thunder, Kaidi Lõo, Antti Arppe, Jordan Lachler, Sjur Moshagen, and Trond Trosterud. 2014. Modeling the Noun Morphology of Plains Cree. *Proceedings of the 2014 Workshop on the Use of Computational Methods in the Study of Endangered Languages*. Ed. by Jeff Good, Julia Hirschberg, and Owen Rambow. Baltimore, Maryland, USA: Association for Computational Linguistics. 34–42.

Violin-Wigent, Anne. 2006. OPTIMALITY THEORY: CONSTRAINT INTERACTION IN GENERATIVE GRAMMAR. *Studies in Second Language Acquisition* 28(1). 160.

Zueva, Anna, Anastasia Kuznetsova, and Francis Tyers. 2020. “A Finite-State Morphological Analyser for Evenki”. English. In: *Proceedings of the Twelfth Language Resources and Evaluation Conference*. Ed. by Nicoletta Calzolari et al. Marseille, France: European Language Resources Association. 2581–2589.

Appendix A: Table of Allomorphs

Class contains the name of the morpheme category (if **Form** is specified with < or >), the subcategory (if **Form** is blank), or the gloss/lexical form of the morpheme. **Form** specifies either the directionality of conditioning of a certain class of morphemes, or a phonological form of a certain morpheme. **Inwards** specifies the segment or set of segments that condition the given allomorph. **Outwards** specifies whether the given allomorph occurs when there is material further from the root: this is specified with a *?*. Each morpheme should have one allomorph that is unspecified for either direction: this is the "elsewhere"/default form. **Notes** contains notes for current and future investigations.

	Class	Form	Inwards	Outwards	Notes
0	VRoot	<			
1	1a				
2	swim	atawv			
3	hang.out	aterakvrie			
4	sew	'nikhu			
5	write	hyatu			
6	buy	hninu			
7	make	uni			
8	fold	hwe'nuni			
9	cook	khuni			
10	ascend	rathv			
11	control	anuhtu			
12	do.best	ateweyv'ton			
13	point.at	ahtsatu			
14	pour/spill	awru			
15	put.down	yv			
16	see	kv			
17	organize	rihwahseruni			
18	close	hnhotu			

19	control	anuhtu
20	knit	riseruni
21	speak	atati
22	tell.story	karatu
23	wash/clean	nohare
24	ask	rihwanutu
25	show	na’tu
26	close.door	hnhotu
27	dress	atsherunyanyu
28	move.around	atoriahnru
29	tear	ratsu
30	take	haw
31	2a	
32	become	atu
33	touch	yena’
34	defend	hnhe’
35	become.happy	atshennuni
36	2b	
37	like	nuhwe’
38	learn	weyvtehta’
39	be.ready	atateweyvnvta’
40	get.scared	htru’
41	recover	yehwvta’
42	remember	ehyahra’
43	step.on	rata’
44	understand	’nikuhrayvta’
45	disappear	ahtu’
46	hire	hnha’
47		
48	Reflexive	>

49	Reflexive			
50	REFL	atat(v)		
51	SRFL	ar	a	only some
52		an	i	
53		at		
54		atv		only some
55		a		only some
56				
57	PronPF	>		
58	agent			
59	1sgA	k		palatalizes before y (East)
60	2sgA	ts	[i | y]	
61		hs		first phase only?
62	MsgA	hr	[e | v | o | u]	
63	ha			
64	FsgA	iak	[e | v | o | u]	
65		iawa	a	
66		ie		
67		w	[e | v]	io?
68		wa	a	
69		y	[o | u]	
70		ka		
71	1incl.duA	ty	a	
72		tni		e epenthesis
73	1excl.duA	yaty	a	
74		iakni		e epenthesis
75	2duA	ts	a	
76		sni		e epenthesis
77	MduA	hy	a	

78		hni		
79	FZduA	ky	a	
80		kni		
81	1incl.plA	ty	o	on??
82		twa		
83	1excl.plA	yaky	o	on?
84		yakwa		
85	2plA	ts	o	on?
86		swa		
87	MplA	hun	[i | e | v | o | u]	
88		hawa	a	awa?
89		hati		
90	FZplA	kun	[i | e | v | o | u]	
91		ku	a	
92		kunti		awa?
93	patient			
94	1sgP	wak		
95	2sgP	s	[e | v | o | u]	
96		sa		
97	MsgP	haw	[e | v | o | u]	
98		ho		
99	FsgP	yakaw	[e | v | o | u]	
100		yako		
101	NP	yaw	[e | v | o | u]	
102		yo		
103	1duP	yuki	a	
104		yukni		
105	2duP	ts	a	
106		sni		

107	MduP	hon	[a | e | v | o | u]	
108		hoti		
109	FZduP	yon	[a | e | v | o | u]	
110		yoti		
111	1plP	yuky	o	
112		yukwa		
113	2plP	ts	o	
114		sewa		
115	MplP	hon	[a | e | v | o | u]	
116		hoti		
117	FZplP	yon	[a | e | v | o | u]	
118		yoti		
119	transitive			
120	1sg>2sg	kuy	[a | i | e | v | o | u]	
121		ku		
122	1sg>2du	ky	a	
123		kni		
124	1sg>2pl	ky	o	
125		kwa		
126	1sg>Msg	hiy		
127	1sg>FI	khey		
128	1incl.du>Msg	tshity	a	initial e
129		tshitni		
130	1incl.pl>Msg	tshity	e	
131		tshitwa		
132	1excl.du>Msg	hshaky	a	
133		hshakni		
134	1excl.pl>Msg	hshaky	e	
135		hshakwa		

136	1incl>FI	yethiy			
137	1excl>FI	yakhiy			
138	2sg>1sg	hskw	a	?	
139		takw	a		
140		hsk		?	
141		tak			
142	2sg>1du	hsky	a	?	
143		taky	a		
144		hskni		?	
145		takni			
146	2sg>1pl	hsky	e	?	
148		hskwa		?	
149		takwa			
150	2sg>M	tsh			initial e
151	2sg>FI	hshey			
152	FIsg>1sg	yukw	a		
153		yuk			
154	FIsg>1du	yukhiy			
155	FIsg>2sg	yesa			
156	FIsg>2du	yesthsiy			
157	FIsg>Msg	ruway	o		
158		ruwa			
159	FIsg>FI	yutat			
160	Fisg>FZsg	kuway	o		
161		kuway			
162	Msg>1sg	hakw	a		
163		hak			
164	Msg>1du	hshuky	a		
165		hshuka	i		

166		hshukni		
167	Msg>1pl	hshuky	o	
168		hshukwa		
169	Msg>2sg	hya		
170	Msg>2du	tshits	a	initial e
171		tshitsni		
172	Msg>2pl	tshits	o	
173		tshitsw		
174	Msg>Msg	haw	[e | v | o | u]	
175		ho		
176	Msg>FI	hshakaw	[e | v | o | u]	
177		hshako		
178	Mpl>FI	hshakon	[a | e | v | o | u]	
179		hshakoty		
180	Fpl>Mpl	ruwan	[a | e | v | o | u]	
181		ruwaty		
182	Fpl>Fpl	kuwan	[a | e | v | o | u]	
183		kuwaty		
184	Fpl>FI	yakon	[a | e | v | o | u]	
185		yakoty		
186				
187	Aspect	<		
188	Punctual			
189	PUNC1a	’		
190	PUNC2a	’		
191	PUNC2b	n’		
192	Habitual			
193	HAB1a	s		
194	HAB2a	s		

195	HAB2b	s			
196	Stative				
197	STAT1a	0			
198	STAT2a	’u			
199	STAT2b	’u			
200					
201	Modal	>			
202	Modal				
203	OPT	a(y)e	[s | t]		
204		aa			
205	FUT	v			
206	DUP””FACT	wa’ti	[s | t]		
207		wa’t			
208	FACT	e	[s | t]	?	
209		a		?	
210		wa	w		
211		we	[s | t]		check NA
212		wa’			y disappears
213	Imperative	>			
214	Imperative				
215	IMP	0			
216					
217	PrePron1	>			
218	PrePron1				
219	DUP	t	[a | e | i | o | v | u]		
220		te			

Appendix B: List of Allomorphy Classes

Per Michelson, 1988, natural classes are formulated with the assumption that phono logically, /ʌ̃/ is a mid front vowel and /ũ/ is a mid back rounded vowels: essentially, nasalized versions of /e/ and /o/.

Segment Set	Natural Class
{/a/}	/a/
{/i/}	/i/
{/e/}	/e/
{/o/}	/o/
{/i/,/j/}	Palatal Sonorants
{/e/,/ʌ̃/}	Mid Front Vowels
{/o/,/ũ/}	Mid Back Vowels
{/e/,/ʌ̃/,/o/,/ũ/}	Mid Vowels
{/i/,/e/,/ʌ̃/,/o/,/ũ/}	Non-Low Vowels
{/a/,/e/,/ʌ̃/,/o/,/ũ/}	Non-High Vowels
{/a/,/i/,/e/,/ʌ̃/,/o/,/ũ/}	Vowels
{/w/}	/w/
{/s/,/t/}	Coronal Obstruents