# A finite-state model of Innu verbal inflection

**Loïc Daignault-Pichette**
OLST, Université de Montréal
loic.daignault-pichette@umontreal.ca

**François Lareau**
OLST, Université de Montréal
francois.lareau@umontreal.ca

## Abstract

This article describes a finite-state transducer (FST) of Innu verbal inflection. Unlike approaches that model word formation by concatenating unanalysed inflectional chunks to a root, the proposed model is linguistically motivated and reflects the largely agglutinative nature of Innu verbal morphology. Our implementation relies on a sequence of intermediate representations, making it modular. It covers 19,628 unique forms for 68 lemmas and it has been evaluated against data from the most comprehensive resources currently available. We argue that FSTs can serve not only as a practical natural language processing tool, but also as a means of testing the descriptive adequacy of a morphological analysis. We discuss the challenges encountered in developing this model.

## 1 Introduction

Finite-state transducers (FSTs) have been used for morphology in natural laage processing (NLP) since the mid-1980s and have been applied to a wide variety of languages (Karttunen and Beesley, 2005). A morphological FST represents a set of pairs of symbol strings mapping a *deep level* to a *surface level*,[1] either for synthesis or analysis. Analysis maps a word to its valid analyses. Synthesis is the inverse: it maps an abstract morphological representation to the corresponding word. These mappings can be one-to-many, thus modelling ambiguity and homography. A FST can also be made productive, meaning it can analyze or synthesize new forms, independently of its internal lexicon. We have set out to create an open-source morphological model of this type to help develop computational resources for Innu.[2] The value of this model lies in its subject and its structure.

**The subject.** The model targets Innu, an Algonquian language spoken in the eastern part of the Québec-Labrador peninsula with approximately 10,000 speakers (Tanner, 2025). Models have been created for other Algonquian languages (Hammerly et al., 2026; Bowers et al., 2017; Harrigan et al., 2017; Davis et al., 2021; Kazeminejad et al., 2017) and for other Indigenous languages of North America (Arppe et al., 2016; Forbes et al., 2021; Lachler et al., 2018), but to our knowledge, only two automatic verb conjugators have been developed for Innu—one of which has not been published—(Junker and Stewart, 2008), and none of them are still in use. Our model is therefore a valuable resource for the language processing of Innu.

The model presented here only deals with inflectional morphology, not with derivational morphology. In Innu, as in many other languages, derivational morphology is much more idiosyncratic and therefore difficult to generalize and implement. Also, although some of it is suitable for implementation, much of it is either not well documented or sensitive to factors such as vowel length, which are not accounted for in the linguistic resources currently available.

**The structure.** The structure of the model was designed to simplify its implementation from a theoretically motivated linguistic description. Computational models are often shaped by practical considerations such as ease of implementation, operational efficiency, or compatibility with available tools, rather than by theoretical considerations such as the plausibility of the description, its logical elegance, or its consistency with descriptions of other languages. In our implementation, we have paid close attention to theoretical considerations.

The expressiveness of automata theory and the set of high-level operations available in the XFST formalism (Beesley and Karttunen, 2003),[3] allow

[1] Often referred to as *upper level* and *lower level*, respectively.

[2] The source code is available at `https://codeberg.org/ldp/innu-aimuna`, under GPLv3. However, the test corpus we used is not included, as it is copyrighted.

[3] We use Foma (Hulden, 2009), a free clone of XFST.



*Proceedings of the International Conference LAnguage TEchnologies for Low-resource Languages*, pages 80–88
28 – 30 September 2026, Fes, Morocco ©LaTeLL 2026
https://doi.org/10.26615/978-954-452-112-7_010

for the faithful representation of a wide range of morphological phenomena and can be used to describe morphological transducers in a manner that resembles certain classical formal linguistic descriptions. This allows us to envisage that morphological descriptions can be systematically tested against a large amount of real data, and thus subjected to the quantitative comparison of their descriptive adequacy.

Since the 1990s, rule-based approaches have gradually given way to machine learning methods. In particular, deep learning has led to models that demonstrate real competence in NLP, including for morphology (Baxi and Bhatt, 2024). Our rule-based approach is motivated by theoretical, practical, and social considerations. First, since deep learning models are less transparent, they would be less suitable for testing linguistic descriptions. Second, the corpus of available texts in the Innu language is too small to train a statistical or neural model for morphological processing. In addition, some ethical issues, such as those related to the ownership of the data used for training, are particularly sensitive when it comes to Indigenous peoples (Junker, 2024). Finally, it is well known that statistical or neural models trained on small corpora make many errors, which poses a problem when attempting to describe a language for which very few reference works are available to speakers. Indeed, due to deficiencies in the transmission of the Innu language and the fact that Innu people are primarily educated in French or English, many speakers suffer from linguistic insecurity in both spoken and written forms. To avoid exacerbating this further, it is important to provide reliable resources. In this regard, since they are deterministic, rule-based models offer greater guarantees of predictable behaviour than machine learning models, at the cost of more limited coverage.

## 2 Innu verbal inflection

To make the description of the model easier to understand, this section provides a brief overview of Innu verb inflection. This overview is based on a morphological analysis that we conducted specifically for the purposes of the model. For this reason, the analysis is based on standardized spelling rather than on the spoken language, and we considered the meaning of linguistic units only insofar as it could inform the study of their form. Furthermore, it largely follows established Algonquianist terminology. Our analysis proved to be similar in many respects to that of Agnès (2023), of which we only became aware once our work was already well underway. Both adopt a positional approach to Innu's inflectional morphology. However, when confronted to forms that could be analyzed but only imperfectly with more affixes than less, we favoured the more radically agglutinative one, whereas Agnès opted in these cases for a slightly more fusional analysis. For example, the segment *nan* in *nuapama**nan**at* ('we.EXCL see them') is one morpheme that encodes the local persons as well as the first person plural in Agnès's analysis and two morphemes in ours, where *n* is responsible for the local persons and *an* for the first person plural. This difference of approach leads to a small difference in the number of positions we recognized. It should be noted that Agnès paid a much closer attention to meaning than we did.

Innu verbs are inflected for person, number, obviation, order, tense, and mood. The person category distinguishes the three usual persons. There are two numbers: singular and plural. The first-person plural is also marked for the inclusion (or exclusion) of the addressee. Obviation is a category characteristic of Algonquian languages that ranks third persons according to their peripheral nature (Volchok, 2019). It opposes the *proximate* (central) and the *obviative* (peripheral), and some paradigms allow up to four degrees of obviation, including the proximate. In traditional terminology, these degrees are often conflated with the person category, where the third person corresponds to the proximate, the fourth to the first degree of the obviative, the fifth to its second degree, and the sixth to its third degree. Since the categories of person and obviation are semantically and morphologically distinct, in our model we have chosen to distinguish them. The three orders are the independent, the conjunct, and the imperative. The tenses are present, past and future. There are numerous moods, and many have an evidential meaning. Not all combinations of order, tense, and mood are permitted. We have excluded from our model inflections that are expressed by an auxiliary, such as the future tense (except in the imperative, where it is synthetic).

From a lexical standpoint, verbs are classified according to their transitivity and to the animacy of their subject (for intransitive verbs) or their object (for transitive verbs—the subject of a transitive verb must be animate). The intersection of

| Form | PPERS | R | DÉC | DIR | OBV | SPERS | NBLC | TIR | NBNLC |
|---|---|---|---|---|---|---|---|---|---|
| 2SG→1SG | tsh | uapam | | | | in | | atak | |
| 2PL→1SG | tsh | uapam | | | | in | au | atak | |
| 1SG→2PL | tsh | uapam | | it | | in | au | atak | |
| 1PL.EX→3SG | n | uapam | | a | | n | an | atak | |
| 3SG→1PL.EX | n | uapam | | iku | | n | an | atak | |
| 3SG→2PL | tsh | uapam | | ik | | u | au | atak | |
| 3PL→2PL | tsh | uapam | | ik | | u | au | atak | at |
| 1PL.IN→3OBV | tsh | uapam | im | a | n | | an | atak | a |

Table 1: Positional analysis of different forms of the VTA *uapameu* ('to see') in the present indirect independent.

these features yields the four commonly used Algonquian verb classes: animate intransitive verb (VAI), inanimate intransitive verb (VII), transitive animate verb (VTA), and transitive inanimate verb (VTI). These classes are important because verbs within the same class share the same morphemic patterns. In most cases, the inflectional and syntactic behaviours of members of the same class align, but when they do not, it is inflection that prevails. For example, the VAI class includes a subclass called VAI+O, whose members require a direct object and are thus actually transitive. Verbs in this subclass, however, conjugate like any other VAI. All verbs take the subject's features, but VIIs only allow for the third-person form (since the first two persons cannot be inanimate), and VTAs additionally take the object's features. The VTA paradigm is defective: it does not allow all combinations of subject and object features. For a given Order-Tense-Mood combination, VIIs have at most 4 different forms, VAIs and VTIs have 8, and VTAs have 45 subject-object combinations.

The verbal morphology of Innu can be fairly well represented by a sequence of *positions*, or *slots*. Depending on the context, a position may be occupied by a morpheme or left empty. Morphemes that appear together in the same form cannot, therefore, occupy the same position. This criterion is decisive for determining the number of positions but insufficient for assigning morphemes to the various positions. Indeed, it often happens that the distribution of a morpheme allows for its assignment to several adjacent positions left empty by other morphemes. In such cases, the semantic criterion, according to which morphemes occupying the same position tend to express the same class of grammatical meanings, allows us to resolve the issue. In this way, we assign to each position the expression of specific features. We have identified 9 positions: the person prefix (PPERS),[4] the root (R), the shift (DÉC), which in VTAs is used to "shift" a third person by one degree of obviation, the directional (DIR), the obviative (OBV), the person suffix (SPERS), the local person number (NBLC), the tense or mood (TIR), the non-local person number (NBNLC), and the perceptive (PERCEP), that is only used in the perceptive mood. Not all verb classes realize the same positions, and only VTAs realize all of them. Table 1 presents the positional analysis of different forms of the VTA *uapameu* ('to see') in the present indirect independent.

Verbal roots can also be decomposed in the same way, but since they belong to derivational morphology and the lexical resources we use represent them as simple roots, we treated them as unanalyzable.

The DIR position deserves further explanation. It is found in VTAs and primarily distinguishes between *direct* forms and *inverse* forms. Its role is better understood through the following hierarchy:

$$2 > 1 > 3 > 3\text{OBV}$$

The distribution of some morphemes in VTAs, and many other phenomena in Innu, can be explained largely by this hierarchy (Drapeau, 2014). For example, the second-person prefix *tshi–* can contribute, in VTAs, to the expression of the person combinations 2→1, 1→2, 2→3, and 3→2. By default, or when the DIR position contains a direct morpheme, only combinations where the subject is higher than the object in the hierarchy are permitted (2→1 and 2→3); when the DIR position contains an inverse morpheme, the condition is reversed.

Morphemes are not entirely orthogonal. Some-

[4] We use French abbreviations in this paper, to echo the symbols used in our code. Since most native speakers of Innu speak French, more than English, having the code in French increases maintainability.

times a combination of morphemes has an unexpected meaning, and sometimes several combinations could logically express the same meaning, while only one is actually used. In fact, the two major dialects, Eastern and Western, have sometimes retained different combinations to express the same meaning. Also, especially in the conjunct order, some morphemes are fusional, which means that they span multiple positions in our analysis. For example, the morpheme *ishk*, used in the conjunct order, realizes both the direction (inverse) and the person suffix (second and third person) positions.

This positional analysis views Innu morphology as mostly concatenative. This is largely true, but certain moods of the conjunct order exhibit vowel alterations called *initial changes* (IC in the glosses below) that affect the first vowel of a verb's root. These alterations are regular and depend on vowel quality and length. For example, the forms of the iterative subjunctive conjunct differ from those of the proximate subjunctive conjunct only through the application of these alterations (Drapeau, 2014).

(1) Pipu-k-i, uemut ni-ka kushpi-n
be.winter-NLC-SBJV surely 1-FUT go.inland-LC
'When winter comes, I will surely go inland.'

(2) Pepu-k-i, nanitam ni-kushpi-n
be.winter.IC-NLC-SBJV always 1-go.inland-LC
'In winter, I always go inland.'

The form *pepuki* follows the rule where a short *i* becomes a long *e* in a form that presents initial change. However, vowel length is not consistently indicated in dictionaries and cannot be deduced from general rules. Furthermore, dialects diverge on the length of the first vowel of certain roots. This is the case with *tshitshipańu* ('to begin'), whose *i* is long in eastern dialects and short in western ones (Drapeau, 2014). For this reason, the inflected forms of the roots have not been standardized.

We found two patterns that do not fit well into our analysis because they present the apparent duplication of a morpheme, which we discuss here using the VTA *uapameu* ('to see') as an example. In the past indicative independent form 2PL→3OBV, *tshuapamimat**au**aua* is used instead of the expected *tshuapamimataua*, with the apparent duplication of the second-person plural morpheme *au*. Similarly, in the western dialect, the present indicative conjunct 3OBV→1PL.EX form, *uapam**im**imiti*, appears to duplicate the shift morpheme *im*. In both cases, the two identical segments could be treated either as a single special morpheme or as two instances of the same morpheme. Our implementation adopts the latter analysis, although the former would be more compatible with a strictly positional account.

## 3 Implementation

As discussed by Lachler et al. (2018), Muradoglu et al. (2020) and Hammerly et al. (2026), early in the design of a morphological transducer, one has to choose between favouring a chunk-based approach, which treats word formation using unanalysed blocks affixed to a root, a concatenative one, where morphemes are independently concatenated following the most general rules possible in a manner that can be similar to that used in classical morphological descriptions, or a mix of both. In a way, the chunk-based approach treats a language's morphology as if it were completely fusional. Both approaches can be used to model any language's morphology. In agglutinative languages, where the differences are the most salient, there are examples of implementations using chunks (Hammerly et al., 2026) or concatenation (Bowers et al., 2017) (note that both models target Ojibwe dialects).

A chunk-based approach simplifies the treatment of irregularity because it enumerates inflectional forms in relatively large units, thus treating both regular and irregular forms in the same way. By contrast, a concatenative approach requires separate rules for both regular patterns and exceptions, which increases implementation complexity but also improves generalization and linguistic transparency. However, the broader the scope of a rule, the greater the likelihood that it will overgenerate. All things considered, and given our interest in linguistic plausibility and model explainability, we have opted for a concatenative approach.

To address the inherent complexity of describing the morphology of an agglutinative language, we decided not to directly map the deep level to the surface level, but rather to create four intermediate levels with well-defined structures and roles. Each level isolates a distinct descriptive problem: feature ordering, positional structure, morpheme selection, canonical exponent insertion, allomorphic realization, and orthographic normalization. This sequence of levels of analysis is loosely inspired by the Meaning-Text Theory (Mel'čuk, 1993) and is made natural by the composition operation, sup-

ported by any finite transducer. For example, let $T_1$ be a transducer that removes all circumflex accents, which are used in the definition of our model to mark vowel length but are not used in Innu orthography, and $T_2$ a transducer that lowers the superscript of *u* ($^u$) to *u* when it is not at the end of a word, as per spelling rules. It is possible to construct a transducer $T_3$ equivalent to the successive application of $T_1$ and $T_2$.

$$T_1(\text{tûtam}^{\text{u}}\text{at}) = \text{tutam}^{\text{u}}\text{at}$$
$$T_2(\text{tûtam}^{\text{u}}\text{at}) = \text{tûtamuat}$$
$$T_3 = T_1 \circ T_2$$
$$T_3(\text{tûtam}^{\text{u}}\text{at}) = \text{tutamuat}$$

The composition operation makes it possible to construct abstractions and to modularize the implementation without affecting the model's runtime performance. The four intermediate levels are therefore only relevant for the description of the model and are merged at compile-time into one final transducer, which directly maps the deep level to the surface level.

### 3.1 Defining the lexicon

The internal lexicon of the transducer is defined in a file that lists each lemma, along with any necessary lexical information.

```
tshimuan+impers+VII
tau+IdTau+Râ+VAI
ashtueim^u+VTI
uitshieu+VTA
```

The verb class is specified for each verb using the features `+VII`, `+VAI`, `+VTI` and `+VTA`. Some VIIs are impersonal; they are marked with `+impers`, which informs the model that it has to block the realization of all non-third-person singular forms for this verb. Moreover, since it is not always possible to determine the root of a verb based solely on its citation form, we sometimes have to encode information about the root. For instance, in the excerpt above, `+Râ` indicates that the root ends in /aː/ (the final *u* of the citation form *tau* being a third-person suffix).

In the overwhelming majority of cases, knowing the verb class is sufficient to derive the inflectional form from a root. It therefore seems that, by directly specifying roots rather than citation forms, we could do without the features that allow us to recover a verb's root. However, since we want to productively generate forms from a citation form (because all the resources of the language refer to it), there must be a way to provide the information necessary to discover the root. Features such as `+Râ` are therefore unavoidable, as is the infrastructure that makes them work. It then seemed simpler to us to use the same mechanism to define the internal lexicon as to specify the productive lexicon.

The uniformity of the definition of the internal lexicon and of the productive one goes even further. *Tau* ('to be (somewhere)') belongs to a small group of irregular verbs in Innu. It would have been possible to account for its irregularity using rules targeting the string *tau* (or *tâ*, in cases where it occurred after the recovery of the roots). Instead, we decided that these rules would target a symbol specific to the lemma, `+IdTau`. This method has three advantages. First, it makes the irregularity of verbs bearing such features explicit. Second, it opens the door to the productive processing of irregular verbs and some of their derivatives that would not be in the internal lexicon. Finally, since productive forms are generated in the same way as non-productive ones, by verifying the correctness of the non-productive forms, we also verify a portion of the productive forms.

Toward the end of the transducer's compilation, features such as `+Râ` and `+IdTau` are removed from the deep level of non-productive forms, so that knowledge of the idiosyncrasies of a verb is not required to synthesize its non-productive form. Those features need to be provided explicitly to synthesize productive forms, as otherwise the model would have no way to recover roots with certainty, which is necessary to produce correct forms. However, for cases where the root can be deducted unambiguously from the citation form, the user is not required to provide a root extraction feature.

### 3.2 Levels

**Deep level.** The deep level is the starting point of synthesis and the endpoint of analysis. Its elements are divided into three classes: lemma, lexical features, and inflectional features. They are ordered as follows: the citation form must precede the lexical features, which must precede the inflectional features. Within a class, the elements must also follow a specific order. The order of lexical features is: `+Prod` (which triggers productive mode), root finding feature, `+impers`, length of the first vowel, and verb class; the order of inflectional features

| Level | Structure |
|---|---|
| Deep | nipau+Râ+VAI+Est+Indép+Ind+Prés+1+Pl+Ex |
| Positional | [PPers]nipâ[Déc][Dir][Obv][SPers][NbLc][Tir][NbNLc][Percep]+VAI+Râ+Est+Indép+Ind+Prés+1+Pl+Ex |
| Morphological | <1>nipâ<Radical><Lc><1Pl>+VAI+Râ+Est |
| Morphemic | nit<1>nipâ<Radical>n<Lc>an<1Pl>+VAI+Râ+Est |
| Morphic | ni<1>nipâ<Radical>n<Lc>an<1Pl>+VAI+Râ+Est |
| Surface | ninipanan |

Table 2: All levels of representation for *ninipanan*, the first-person exclusive plural of the independent present indicative of the VAI *nipau* ('to stand').

is: order, mood, tense, person, number, inclusivity, and obviation. This strict ordering is a consequence of the choice of FSTs for implementation.

The two main dialects of the Innu language are distinguished in the deep level by two mandatory features: `+Est` for the Eastern dialect and `+Ouest` for the Western dialect. Most often, they map to the same surface form, but they are required as the two dialects don't share all forms. In general, a deep form maps to only one surface form, but one surface form can map to many deep ones.

To indicate the length of the first vowel of a root, two features are used: `+VICourte` (for a short vowel) and `+VILongue` (for a long vowel). They are used for the application of initial change. As this length is not well described nor consistent across dialects, these features cannot be set in the lexicon but have to be supplied explicitly at runtime to conjugate a form that triggers initial change. In fact, for these forms, the model will generate nothing unless one of them is given. We had to resort to this mechanism because of the absence of information on vowel length in available resources.

For non-productive forms, lexical features are limited to the verb class and the dialect; for productive forms, they also include the features necessary to identify the root, as well as the features `+impers`, `+Prod`, etc.[5] Inflectional features specify the order (`+Indép` for *independent*, `+Conj` for *conjunct*, or `+Imp` for *imperative*), the mood (`+Ind` for *indicative*, `+Indir` for *indirect*, etc.), tense (`+Prés` or `+Pas`), person (`+2`, `+1`, or `+3`), number (`+Sing` or `+Pl`), inclusivity (`+In` or `+Ex`) and obviation (`+Ntr`, `+Obv`, `+Obv2` or `+Obv3`).

[5] As mentioned above, non-productive forms bear the same features as productive forms (except `+Prod`) until very late in the model's compilation. This explanation concerns runtime.

**Positional level.** At this stage, roots are identified using features such as `+Râ` or via default rules. This representation introduces the positions we discussed earlier. Lexical and inflectional features are preserved and moved to the right.

```
{a} -> {â} || _ \Lettre* "+Râ"
```

***Sample rule.*** *Replace `a` with `â` when only non-letters separate it from `+Râ`.*

**Morphological level.** At this level, a position is either filled with a symbol representing a morpheme or left permanently empty. For example, at the morphological level in table 2, the position `[PPers]` is filled with the morpheme `<1>` (first person), `[SPers]` with `<Lc>` (local person) and `[NbLc]` with `<1Pl>` (first person plural). `<Radical>` is the root, which can be treated like any other morpheme. These morphemes are selected based on lexical and inflectional features. All the information necessary for the realization of the morphemes must be conveyed either by the identity of these morphemes or by the lexical features inherited from the deep level, since inflectional features are not carried over to the morphemic level.

```
"[PPers]" -> "<1>" || \
    _ [$["+1"] - $["+In"]] \
    & $["+Indép"]
```

***Sample rule.*** *Fill `[PPers]` position with morpheme `<1>` in non-inclusive independent first-person forms.*

**Morphemic level.** Here, the canonical forms of morphemes are placed to the left of the morpheme symbol. These symbols are retained because they uniquely identify the morphemes and can therefore resolve ambiguities between homographic forms that would otherwise not be targeted by the same morphophonological rules. For this reason, we

generally prefer to work with symbols rather than character strings, and we delay the introduction of strings until as late as possible.

```
"<1>" -> {nit} "<1>"
```

***Sample rule.*** *Add the canonical form of the first person morpheme symbol before its symbol.*

**Morphic level.** The morphic level is similar to the morphemic level, but it uses fully realized morphemes rather than canonical forms.

```
{t} -> [] || \
     _ ["<2>"|"<1>"] \Voy
```

***Sample rule.*** *Remove the final* t *of the <2> or <1> morphemes before a non-vowel.*

**Surface level.** The surface representation is a grammatically correct inflected form of a verb. Compared to the morphic level, morpheme symbols and lexical features no longer appear. In addition, some general spelling rules are applied. For example, the superscript *u* (ᵘ) is lowered *(u)* when it is not at the end of a word; the acute accent on *ń* is removed for eastern dialects; and the circumflex accent indicating a long vowel is removed, since standard spelling does not represent vowel length.

### 3.3 Coverage

The non-productive subset of the model currently contains 68 roots and generates 113,682 forms, of which 19,628 are distinct. This discrepancy is largely artificial and has to do with the structure of the deep level. Indeed, the presence of a feature marking the dialect (`+East` or `+West`) results in nearly doubling the number of feature-form mappings without adding many distinct forms. This is also true for `+VICourte` and `+VILongue`, which are only useful with initial change. Also, some systematic syncretisms are not treated as such at the deep level. For example, the plural obviative marker is systematically conflated in the language with the proximate plural marker for inanimate nouns and with the singular obviative marker for animate nouns. We chose to distinguish these forms at the deep level for simplicity, but one could argue that this distinction simply does not exist even at a semantic level. Overall, only a small portion of this discrepancy is attributable to actual homography.

Throughout the model's development, we tested it against the forms contained in the conjugation tables available on `verbe.innu-aimun.ca`, as well as in the grammar by Drapeau (2014). These resources contain 12,068 distinct forms (27,070 in total)[6] distributed across 68 lemmas and thus contain 61% of the 19,628 distinct forms produced by the model (24% of all 113,682 forms). Our system is able to generate every form in this evaluation set, fully covering the available reference paradigms. As for the generated forms that are not in the evaluation set, they are not necessarily false positives. Indeed, the reference material does not exhaustively list all correct forms for the 68 lemmas they contain. For example, the VTA *natutueu* is only listed for the imperative mood in the resources we employed. It thus contributes 1,808 forms to the model but only 72 to the test corpus. For the other moods, *petueu* is used to exemplify its paradigm. Upon closer inspection, we confirmed that the generalizations encoded in our transducer are consistent with the available paradigmatic reference data.

## 4 Discussion

Since our model is rule-based, we were able to refine it iteratively until it covered all forms in the available reference dataset. This result is meaningful because the conjugation tables from which the evaluation set is drawn were designed specifically to present and summarize the diversity of the language's regular verbs. It should nevertheless be interpreted as coverage of the attested reference paradigms rather than as an absolute measure of language-wide accuracy. As for irregular verbs, we have implemented only one, *tau* ('to be (somewhere)'). Thus, we know that our infrastructure supports them, but many have not yet been described by linguists, making it impossible to integrate them into the model.

Among North American Indigenous languages, Innu is relatively well documented: it has a grammar spanning several hundred pages (Drapeau, 2014), a substantial and actively maintained online dictionary (`dictionnaire.innu-aimun.ca`), and a widely adopted standard orthography. Nevertheless, some aspects of the language remain less thoroughly documented. Thus, our task was made somewhat more difficult by the lack of standardization regarding changed forms and by the fact that the length of the first vowel is not systematically specified. Similarly, difficulties arise from the fact

[6] We temporarily excluded 123 forms from our test corpus, as the two resources do not agree on them. We are in the process of checking if either of them contains errors, or if we need to include the forms from both.

that the conjugation of irregular verbs has not been officially established, including for some very common verbs such as *tau* ('to be (somewhere)'). This verb is, in fact, the only irregular verb included in our model, and we were able to include it only because its forms were specifically provided to us.

We also aimed to implement a linguistically plausible description. Of course, our analysis of Innu verbal inflection has several limitations. In particular, being based exclusively on written data, it ignores vowel length, which standard orthography does not capture. Furthermore, the two exceptions to our positional analysis suggest that some parts of the verbal paradigms are more diachronically relevant than synchronically. It should be noted, however, that our analysis was conducted primarily for the purposes of this model, and we consider that it proved adequate for this purpose. It contributes to showing that a theoretically motivated model is viable even for a morphologically rich language such as Innu. Also, we showed that the use of multiple clearly defined intermediate levels of description can be very valuable to manage the inherent complexity of an agglutinative language.

We should point out that this project, although inspired by the many high-level operations offered by the XFST formalism, has revealed the limitations of this formalism. In particular, the limited means of abstraction it offers and its relatively strict adherence to finite-state calculus at compile-time led us to hope for a formalism that combines a task-tailored syntax with more powerful means of abstraction, allowing for the use of Turing machines at compile-time. For example, despite the obvious operationalization of the person hierarchy presented above, we were forced to define a large part of its implications by enumeration.

Finally, one of the main challenges in developing digital resources for the Indigenous languages of North America is the lack of resources—both in terms of funding and personnel. We do not yet know what impact our approach will have on this issue. On the one hand, our model is undoubtedly more complex than it would have been had we used a fusional approach. On the other hand, the abstractions on which it is built allow it to be modified (by adding a root, a morpheme, etc.) in a much more localized and productive manner. This modularity should make the model easier to maintain and extend, provided that its implementation is accompanied by clear documentation.

## 5 Conclusion

We presented a FST for the synthesis and analysis of Innu verbal inflection. This model departs from chunk-based implementations by encoding a linguistically motivated concatenative analysis of Innu verbal morphology. Its architecture relies on positional structure and on a sequence of intermediate representations that separate feature organization, morpheme selection, allomorphic realization, and orthographic normalization.

The current implementation covers the regular verbal paradigms represented in the available reference materials and includes one irregular verb, *tau* ('to be (somewhere)'). Remaining irregular verbs require further description by linguists before they can be fully incorporated.

A morphological transducer lends itself to an array of practical applications. It enables, for instance, the dynamic generation of conjugation tables. Current practical resources on Innu verb conjugation present it using a small set of model verbs, but Innu's verbal inflection is very rich, features many paradigms, and includes irregular verbs, so it would be useful to have an interface that allows users to look up the inflected forms of any verb in the language. We are currently working on such an interface. Also, the rich inflectional system of Innu often makes it difficult for a learner to isolate a verb's root and, consequently, to figure out its lookup form in a dictionary. Adding a morphological analysis engine to an electronic dictionary would allow users to access an entry via any of its inflected forms. Our model could eventually be expanded to include forms from other parts of speech, which could serve as the basis for a lemmatizer or a spell checker. Finally, Innu, like many other Indigenous languages, has very few educational resources. Morphological analysis and synthesis could play an important role in the creation of interactive educational exercises.

Beyond its immediate applications, the model shows that finite-state implementation can function as an empirical testbed for evaluating the descriptive adequacy of morphological analyses in a low-resource language context.

## Acknowledgements

We would like to thank Yvette Mollen for her invaluable feedback and Marie-Odile Junker for sharing lexical data with us.

## References


Sauvane Agnès. 2023. *La morphosyntaxe algonquienne: vers un nouveau modèle de description à partir des données de l'innu-aimun, langue algonquienne du Québec*. Ph.D. thesis, Sorbonne Université.

Antti Arppe, Christopher Cox, Mans Hulden, Jordan Lachler, Sjur N. Moshagen, Miikka Silfverberg, and Trond Trosterud. 2016. Computational modeling of verbs in dene languages: The case of tsuut'ina. In *Working Papers in Dene Languages 2016*, pages 51–69. Alaska Native Language Center.

Jatayu Baxi and Brijesh S. Bhatt. 2024. Recent advancements in computational morphology: A comprehensive survey. *ArXiv*, abs/2406.05424.

K.R. Beesley and L. Karttunen. 2003. *Finite State Morphology*. CSLI studies in computational linguistics. CSLI Publications, Stanford.

Dustin Bowers, Antti Arppe, Jordan Lachler, Sjur Moshagen, and Trond Trosterud. 2017. A morphological parser for odawa. In *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 1–9, Honolulu. ACL.

Fineen Davis, Eddie A. Santos, and Heather Souter. 2021. On the computational modelling of Michif verbal morphology. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2631–2636, Online. ACL.

Lynn Drapeau. 2014. *Grammaire de la langue innue*. Presses de l'Université du Québec, Québec.

Clarissa Forbes, Garrett Nicolai, and Miikka Silfverberg. 2021. An FST morphological analyzer for the gitksan language. In *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 188–197, Online. ACL.

Christopher Hammerly, Nora Livesay, Antti Arppe, Anna Stacey, and Miikka Silfverberg. 2026. Ojibwemorph: an approachable finite-state transducer for ojibwe (and beyond). *Lang Resources & Evaluation*, 60(27).

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

Mans Hulden. 2009. Foma: a finite-state compiler and library. In *Proceedings of the Demonstrations Session at EACL 2009*, pages 29–32, Athens, Greece. ACL.

Marie-Odile Junker. 2024. Data-mining and extraction: the gold rush of AI on indigenous languages. In *Proceedings of the Seventh Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 52–57, St. Julians, Malta. ACL.

Marie-Odile Junker and Terry Stewart. 2008. Building search engines for algonquian languages. *Algonquian Papers - Archive*, 39.

Lauri Karttunen and Kenneth R. Beesley. 2005. Twenty-five years of finite-state morphology. In Antti Arppe, Lauri Carlson, Krister Lindén, Jussi Piitulainen, Mickael Suominen, Martti Vainio, Hanna-Maria Westerlund, Anssi Yli-Jyrä, Juho Tupakka, and Björn Gamb, editors, *Inquiries into words, constraints and contexts: Festschrift in the honour of Kimmo Koskenniemi on his 60th birthday*, pages 71–83. CSLI Publications, Stanford.

Ghazaleh Kazeminejad, Andrew Cowell, and Mans Hulden. 2017. Creating lexical resources for polysynthetic languages—the case of Arapaho. In *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 10–18, Honolulu. ACL.

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

Igor Mel'čuk. 1993. *Cours de morphologie générale*, volume 1. Les Presses de l'Université de Montréal/ CNRS Éditions, Montréal/Paris.

Saliha Muradoglu, Nicholas Evans, and Hanna Suominen. 2020. To compress or not to compress? a finite-state approach to Nen verbal morphology. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 207–213, Online. ACL.

Adrian Tanner. 2025. Innu (montagnais-naskapi). L'Encyclopédie canadienne.

Irina Volchok. 2019. A synthesis of obviation in algonquian languages. Master's thesis, University of Manitoba.