On the Computational Modelling of Michif Verbal Morphology

Fineen Davis, Eddie A. Santos
National Research Council Canada
{firstname.lastname}
@nrc-cnrc.gc.ca

Heather Souter
Prairies to Woodlands Indigenous
Language Revitalization Circle
p2wilrc@gmail.com

Abstract

This paper presents a finite-state computational model of the verbal morphology of Michif. Michif, the official language of the Métis peoples, is a uniquely mixed language with Algonquian and French origins. It is spoken across the Métis homelands in what is now called Canada and the United States, but it is highly endangered with less than 100 speakers. The verbal morphology is remarkably complex, as the already polysynthetic Algonquian patterns are combined with French elements and unique morpho-phonological interactions. The model presented in this paper, LI VERB KAA-OOSHITAHK DI MICHIF handles this complexity by using a series of composed finite-state transducers to model the concatenative morphology and phonological rule alternations that are unique to Michif. Such a rule-based approach is necessary as there is insufficient language data for an approach that uses machine learning. A language model such as LI VERB KAA-OOSHITAHK DI MICHIF furthers the goals of Indigenous computational linguistics in Canada while also supporting the creation of tools for documentation, education, and revitalization that are desired by the Métis community.

1 Introduction

In recent years there has been an increase in computational linguistic analysis of Indigenous languages spoken in what is now called Canada, and in particular Algonquian languages such as East Cree (Arppe et al., 2017a), Plains Cree (Harrigan et al., 2017), and Odawa (Bowers et al., 2017). This paper adds Michif—a mixed language of Cree and French origin—to the list with a description of LI VERB KAA-OOSHITAHK DI MICHIF. LI VERB KAA-OOSHITAHK DI MICHIF, which translates as "The Michif verb maker", is a computational model of the verbal morphology of Michif, implemented in the XFST framework of Beesley and Karttunen (2003).

2 Motivation

During the 19th century, marriage between French fur traders and Cree and Anishinaabe women in the Métis homeland was common, and their descendants became known as the Métis peoples (Bakker, 1997; Rosen and Souter, 2009). Michif (ISO 639-3: crg), an Algonquian language, emerged as a mixed language which combined elements of French with the Indigenous Algonquian languages Cree and Saulteaux—a distinct dialect of Ojibwa (Bakker, 1997; Rosen and Souter, 2009). There are many varieties of the Michif language, however LI VERB KAA-OOSHITAHK DI MICHIF is based on a variety spoken mainly in Manitoba, Southern Saskatchewan, North Dakota, and Montana.

Lack of linguistic documentation and analysis restricts the ability to create formalized teaching tools and technologies and hinders the efforts of language learners and activists in Indigenous communities. The number of speakers has decreased significantly. It is difficult to estimate true numbers of speakers of the "intertwined" language (Bakker, 1997), but Michif language activists estimate approximately 50-100 speakers with only a handful presently robust enough to be involved in revitalization work (Souter, 2020).

LI VERB KAA-OOSHITAHK DI MICHIF aims to create a complete model of the verbal morphology of Michif based on the current language data. The output of such a model has many applications, including a smartphone application for conjugating verbs in Michif.[1] LI VERB KAA-OOSHITAHK DI MICHIF represents a collabora-

[1] As of this writing, one such app is in development that incorporates LI VERB KAA-OOSHITAHK DI MICHIF.

Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, pages 2631–2636
April 19 - 23, 2021. ©2021 Association for Computational Linguistics

tive effort between linguists, computer scientists, and the Michif speech community.

The source code for LI VERB KAA-OOSHITAHK DI MICHIF is not publicly available to maintain Métis sovereignty of the language data as requested by the community.

3 Michif linguistic background

The unique mixed origins of Michif are reflected in all linguistic domains. The phonological inventory of Michif is sourced from both French and Cree (with the “Saulteaux” or “Chippewa” dialect of Objibwe being a more minor source language). The pronominals are largely French based and verbal constructions are primarily Algonquian-derived (Sammons, 2019; Prichard and Shwayder, 2014; Rosen, 2007). However, while the morphemes and their concatenations are similar to Algonquian languages such as Plains Cree, the morpho-phonological interactions differ significantly from its sources (Sammons, 2019).

Michif is similar to other Algonquian languages in the degree of polysynthetic complexity, yet there are still distinguishing features that make it unique. Verbs can be highly productive by concatenating morphemes containing many categories of syntactic information; Figure 1 provides an illustration of the basic verbal template. The glosses below show an example of a simple verb form and one of a more complex verb form.

(1) *ayaaw*
ayaa-w
IND.PRS.have.VAI-3SG

‘He/she has.’

(2) *ee-ka-kishkeetamiiyit*
ee-ka-kishkeet-amii-yi-t
CONJ-FUT-know.VTI-3>3-OBV1-OBV2

‘As/that he/she (OBV) knows it.’

Michif verbs can be broken down into four inflectional classes by transitivity and animacy. Transitive verbs combine with the animacy of the object to create their inflectional classes (Transitive Animate (VTA) & Transitive Inanimate (VTI)), while intransitive verbs combine with the animacy of the subject (Animate Intransitive (VAI) & Inanimate Intransitive (VII)) (Wolvengrey, 2011; Sammons, 2019). Person marking, obviation, and direction affixes vary according to the inflectional class. There are also two minor classes, Animate Intransitive Transitive (VAIt) (Sammons, 2019) and Animate Intransitive Transitive animate/inanimate (VAIta/i) (Antonov, 2019).

All inflectional classes can be further broken down into the independent and conjunct orders. The independent order marks person and number with long distance agreement between prefixes and suffixes, and is used to express the indicative mode. The conjunct order expresses the indicative and subjunctive modes, and is typically marked with the ‘*ee-*’ prefix, meaning ‘while/as’.

Preverbs provide inflectional and lexical information before the stem of a verb, but occur *after* the person marking prefix in the independent order, or *after* the conjunct marker in the conjunct order. Grammatical preverbs cover tense or relativization, while lexical preverbs are a closed class that add lexical meaning (Sammons, 2019; Rhodes, 2009). Preverbs are optional, and a verb can be modified by multiple preverbs, as seen in the below example. However, while recursion of preverbs is theoretically possible in Michif, such forms are not used by speakers in practice, so this recursion is not modelled by LI VERB KAA-OOSHITAHK DI MICHIF.

(3) *ni-ka-nohtee-maachi-atoshkaanaan*
ni-ka-nohtee-maachi-atoshk-aanaan
1.IND-FUT.DEF-want-begin-work.VAI-1PL.EXCL

‘We (EXCL) will want to begin to work’

VTA verbs in Michif are the only class which has direction, where actions are either direct or inverse depending on a hierarchy of actors as in other Algonquian languages such as Plains Cree (Harrigan et al., 2017). In the direct VTA forms, the ‘ni-/ki-/∅-’ prefix refers to the subject, while the person marking suffix refers to the object. In the inverse VTA forms, the ‘ni-/ki-/∅-’ prefix refers to the object, while the person marking suffix refers to the subject. This can be observed in the difference between the suffixes in the glossed examples below.

(4) *ki-miyeumin*
ki-miyeum-**in**
2.IND-PRS.like.VTA-**DIR**.2SG>1SG

‘You (SG) like me’

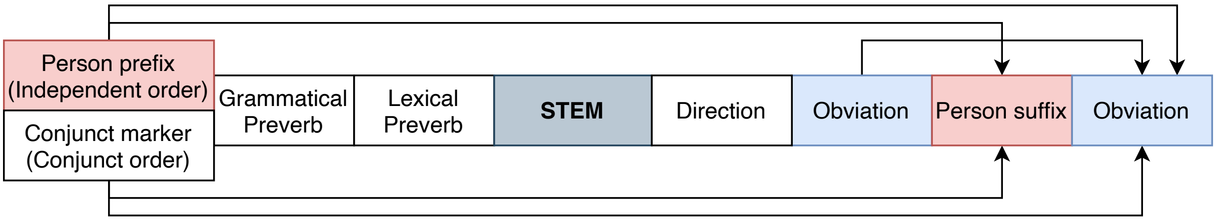

Figure 1: Verbal template of Michif

(5) *ki-miyeumitin*
ki-miyeum-**itin**
1.IND-PRS.like.VTA-**INV**.1SG>2SG

'I like you (SG)'

Obviation occurs across all inflectional classes in Michif and is triggered by obviative nouns (marked or unmarked) (Sammons, 2019). It only occurs in the 3$^{\text{rd}}$ person, and number distinction is not specified, although the verb takes the 3SG suffix[2]. VAI verbs mark obviation in up to two places, relying on long distance dependencies. The '*yi*' morpheme occurs immediately following the stem. The independent order will have an extra '*a*' morpheme following the person marking affix, while the conjunct order will not. This pattern is visible in the examples in the glosses below.

(6) *soñ* *namii*
soñ namii
3SG.MASC.POSS friend
kii-itohteeyiwa
kii-itohtee-**yi**-w-**a**
PST-go.VAI.IND-**OBV1**-3-**OBV2**

'His/her friend (OBV) went'

(7) *soñ* *namii*
soñ namii
3SG.MASC.POSS friend
kii-itohteeyit
kii-itohtee-**yi**-**t**
PST-go.VAI.CONJ-**OBV1**-3

'As his/her friend (OBV) went'

4 Finite-state computational modelling

LI VERB KAA-OOSHITAHK DI MICHIF is a morphological model—a series of composed finite-state transducers (FSTs) called a lexical transducer (Beesley and Karttunen, 2003). There are two primary components of a lexical transducer: the LEXICON and the REWRITERULES. The LEXICON uses a set of labelled sub-LEXICONs which declare the rules for morphological concatenation (Hulden, 2009). This is accomplished by using regular expressions and flag diacritics. Then, the REWRITERULES further constrain the output of the FSTs by using regular expressions to apply phonological restrictions. Foma (Hulden, 2009), a finite-state compiler, takes the LEXICON and the REWRITERULES to create a composed finite-state transducer—in this case, LI VERB KAA-OOSHITAHK DI MICHIF. Using FSTs to model low-resource languages such as Michif is advantageous, as rule-based definitions of verbal paradigms do not rely on access to large, morphologically-tagged corpora which do not exist for Michif.

4.1 Morphological modelling in the LEXICON

The LEXICON allows for the linear concatenation of morphemes in the form of a continuation grammar. Each sub-LEXICON adds two components: morphological tags to the "upper" side, which provide syntactic information, and the surface form morphemes themselves on the "lower" side. Figure 2 illustrates the main sub-LEXICONs of the LI VERB KAA-OOSHITAHK DI MICHIF.

Long distance dependencies are particularly challenging to model with continuation grammars, as they require knowledge of previous states. *Flag diacritics* enable these long distance dependency checks between states so that phenomena such as obviation can be modelled (Hulden, 2009; Beesley and Karttunen, 2003). LI VERB KAA-OOSHITAHK DI MICHIF employs three types of flags: P: **P**ositive set; R: **R**equire feature/value; and D: **D**isallow feature/value. Each feature can be thought of as a set; P flags add a value to this

[2] This paper only focuses on obviation in VAI verbs. See Sammons (2019) for a description of obviation in all verb classes.

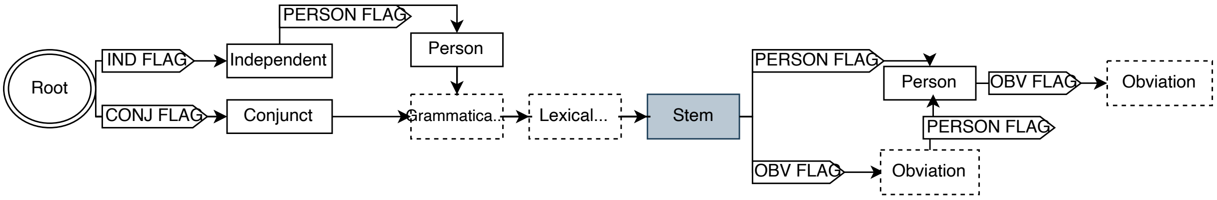

Figure 2: Schematic representation of paths through the LEXICON

feature set, while R flags require a specific value or the entire feature to be set.

As obviation in Michif only occurs with the 3rd person, flag diacritics have to be used to avoid over-application of the obviative concatenation rules. A 'person' feature is used, which distinguishes between 1st and 2nd person forms. When this feature is disallowed, such as with a *@D.person@* flag, only 3rd person forms can pass through. Any forms set to 1st or 2nd person will pass through the null path and continue to the `VAIPersonSuffix` sub-LEXICON, while the 3rd person forms will receive obviative marking before continuing to the `VAIPersonSuffix` sub-LEXICON.

```
LEXICON VAIObviativeMarking
 0                        VAIPersonSuffix;
 @D.person@@P.obv.yi@     VAIYi;

LEXICON VAIYi
 0:yi VAIPersonSuffix;
```

LI VERB KAA-OOSHITAHK DI MICHIF treats obviative as a feature set by a flag. The *@P.obv.yi@* sets the obviative feature, for any forms containing the '*yi*' morpheme. In the example below, the first sub-LEXICON ensures that only obviative forms receive the `[OBV]` tag. The *@D.obv@* flag accepts only those forms which have not previously been set to positive for the obviative feature. The *@R.obv@* flag requires obviative forms, outputting the tag on the "upper" side.

```
LEXICON VAIObviative1
 @D.obv@                 EndVerb;
 @R.obv@[OBV]:@R.obv@    VAIObviative2;

LEXICON VAIObviative2
 @R.order.indep@:@R.order.indep@a EndVerb;
 @D.order.indep@a                 EndVerb;
```

The obviative forms then pass to the second sub-LEXICON, where they are sorted by order using flags. The *@R.order.indep@* requires forms belonging to the independent order, which adds the '*-a*' morpheme. All other forms, i.e. those belonging to the conjunct order, pass through the path with the *@D.order.indep@* flag (disallow any independent forms). Reaching *EndVerb* indicates that the transduction is complete. Table 1 shows examples of lexical outputs before the application of the phonological rules.

The modelling of obviation in Michif takes the opposite approach of the Plains Cree FST in Harrigan et al. (2017). In LI VERB KAA-OOSHITAHK DI MICHIF, the direction and person morpheme are treated as a single unit. Each obviative morpheme is then concatenated individually, rather than as part of the person and direction affixes. For example, LI VERB KAA-OOSHITAHK DI MICHIF would generate '*<eeyi<eew<a>*' instead of '*<eeyieewa>*'. As the intended use of this model is in contexts where all verb forms will only be generated once instead of with each query, the speed of generation is not required to be optimized, and more leniency with regards to stylistic choice is possible.

4.2 Phonological restrictions with REWRITERULES

LI VERB KAA-OOSHITAHK DI MICHIF only accounts for morpho-phonological rules which interact with the verbal morphology of Michif. When morphemes are concatenated, phonemes that would otherwise not co-occur become adjacent, creating the need for phonological rules which handle these issues. The REWRITERULES finite-state machine treats each phonological rule as an individual FST which is then composed to form the final FST (Hulden, 2009; Beesley and Karttunen, 2003).

The *'ni-/ki-'* prefixes host a multitude of morpho-phonological interactions unique to Michif.[3] There is significant vowel reduction as the prefixes are unstressed, which results in the eventual deletion of entire morphemes. Accounting for these alternations requires flag diacritics, such as with obviation, but additionally requires REWRITERULES.

[3] See Bakker (1991) for a full account of morpho-phonological rules in Michif.

<achimostaaweew[VTA][IND][FUT][3][INV][3PL-OBJ][OBV]	<kii-achimostaaw<ikwak>
<ayamihaaw[VAI][IND][PRS][1SG]	<ni-ayamihaa<n>

Table 1: Sample output of Li Verb kaa-Ooshitahk di Michif Lexicon

Lexicon Output	t-Insertion	ni-Deletion
ni-ayamihaan	ni-t-ayamihaan	d-ayamihaan
ni-ka-maachi-kipahaaw		ga-maachi-kipahaaw

Table 2: Sample derivation of the t-Insertion and ni-Deletion RewriteRules

1. t-Insertion
   [..] -> t || [n i | k i] "-" _ Vowel ;

2. ni-Deletion
   n i "-" t (->) d "-" || _ Vowel ;
   n i "-" k (->) g "-" || _ Vowel ;

Following the *'ni-/ki-'* morphemes, [t] is epenthesized when followed by a vowel (Rosen, 2007). Voiceless consonants are voiced when preceded by the *'ni-'* prefix and followed by a vowel. The *'ni-'* prefix is then deleted following the voicing (Bakker, 1991). The RewriteRules combine these two processes into one regular expression rule to avoid over-application of both the voicing and deletion. Table 2 illustrates the application of these rules.

5 Discussion

The latest version of Li Verb kaa-Ooshitahk di Michif includes 105 verb stems, which generate a total of 155,621 verb forms. Compared to the Plains Cree FST, which has around 13,000 verb stems and effectively infinite verbs forms due to recursion, Li Verb kaa-Ooshitahk di Michif is relatively small. However, it is important to recognize that the difference in size is a direct consequence of the linguistic situation of Michif. There is no Gold Standard morphologically-tagged corpus, such as that which exists for Plains Cree (2017b), against which to compare output forms. As a result, the generated forms are hand-verified by speakers (Souter, 2020), so having such a large model is less feasible.

The primary challenge that was faced when creating Li Verb kaa-Ooshitahk di Michif is the lack of consistent language data. The published paradigms conflict in their analysis of Michif verbal morphology, particularly with regards to phonological elements such as vowel length in person marking suffixes (Sammons, 2019; Bakker, 1997; Rosen, 2007). Initially, much of the FST was based on data scraped together from glosses discussing other phenomena, resulting in paradigm gaps. In the absence of accurate language data, the role of speakers becomes vital to ensuring the model generates accurate outputs, beginning with having complete morphological paradigms.

The imperative, subjunctive, and reflexive modes are not currently part of the model. While paradigms for these forms now exist at least partially, in order to implement them the structure of the current FST would have to be built again from the ground up. At the time of creating the current model these paradigms were not available, and so they were not included in the original architectural design and layout of the FST in Foma.

6 Conclusion

Despite the development of Li Verb kaa-Ooshitahk di Michif, much more work is needed to create a complete account of the verbal morphology of Michif. Computational language modelling is an important foundational step towards creating language learning resources and making the creation of those resources more accessible to Indigenous language communities.

References

Anton Antonov. 2019. Loan verb integration in Michif. *Journal of Language Contact*, pages 27–52.

Antti Arppe, Marie-Odile Junker, and Delasie Torkornoo. 2017a. Converting a comprehensive lexical database into a computational model: The case of East Cree verb inflection. In *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 52–56, Honolulu. Association for Computational Linguistics.

Antti Arppe, Katherine Schmirler, Atticus G. Harrigan, and Arok Wolvengrey. 2017b. A morphosyntactically tagged corpus for Plains Cree. In *Papers of the Forty-Ninth Algonquian Conference*, pages 1–16.

Peter Bakker. 1991. The Ojibwa element in Mitchif. *Algonquian Papers-Archive*, 22.

Peter Bakker. 1997. *A language of our own: The genesis of Michif, the mixed Cree-French language of the Canadian Métis*, volume 10. Oxford University Press.

Kenneth R Beesley and Lauri Karttunen. 2003. *Finite State Morphology*. CSLI Publications.

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. Association for Computational Linguistics.

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(4):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. Association for Computational Linguistics.

Hilary Prichard and Kobey Shwayder. 2014. Against a split phonology of Michif. *University of Pennsylvania Working Papers in Linguistics*, 20(1):29.

Richard Rhodes. 2009. The phonological history of Métchif. *Français dun Continent à lautre: Mélanges Offerts à Yves Charles Morin*, 1:423–442.

Nicole Rosen. 2007. *Domains in Michif phonology*. University of Toronto.

Nicole Rosen and Heather Souter. 2009. Language revitalization in a multilingual community: The case of Michif. In *1st International Conference on Language Documentation and Conservation (ICLDC)*, Honolulu.

Olivia N. Sammons. 2019. *Nominal classification in Michif*. Ph.D. thesis, University of Alberta.

Heather Souter. 2020. Personal Communication.

Arok Elessar Wolvengrey. 2011. *Semantic and pragmatic functions in Plains Cree syntax*. Netherlands Graduate School of Linguistics.