OjibweMorph: An approachable finite-state transducer for Ojibwe (and beyond)

Christopher Hammerly[1*], Nora Livesay[2], Antti Arppe[3], Anna Stacey[1], Miikka Silfverberg[4]

[1]Department of Linguistics, University of British Columbia, 2613 West Mall, Vancouver, V6T 1Z4, BC, Canada.
[2]Department of American Indian Studies, University of Minnesota, 150 Pillsbury Drive SE, Minneapolis, 55455, MN, USA.
[3]Department of Linguistics, University of Alberta, 9137 116 St NW, Edmonton, T6G 2E7, AB, Canada.
[4]Independent.

*Corresponding author(s). E-mail(s): chris.hammerly@ubc.ca;
Contributing authors: live0015@umn.edu; arppe@ualberta.ca;
anna.stacey@ubc.ca; mpsilfve@iki.fi;

Abstract

This paper describes the design, evaluation, and application of *OjibweMorph*, a finite-state transducer (FST) for generating and analyzing words in the Central Algonquian language Ojibwe. We created a language-general modular system for creating FSTs from human- and machine-readable spreadsheets, where sets of inflectional and derivational morphology can be defined, combined with a lexical database, and automatically compiled into an FST. We show how this system is applied to generate and analyze the complex nominal and verbal morphology in Ojibwe, with an eye towards how our framework and toolkit can be used to create FSTs for other morphologically complex languages. We evaluate the Ojibwe version of the system by checking the model's performance against a set of inflectional forms and example sentences from the Ojibwe People's Dictionary, and describe the application of the FST to create a linguistically analyzed corpus, an automatic verb conjugation tool for education, a spell-checker, and intelligent dictionary search.

Keywords: morphology, Algonquian, finite-state, Indigenous languages

Accepted for publication in *Language Resources and Evaluation*
October 2, 2025

1 Introduction

This paper describes the structure and evaluation of *OjibweMorph*, a finite-state transducer (FST) for generating and analyzing all possible words in Ojibwe. Ojibwe (ISO 639-3: `oji`) is a North American Indigenous language in the Central branch of the Algonquian family, spoken by close to 30,000 people (Statistics Canada, 2021) in the land area surrounding the Great Lakes in both Canada and the United States. As a polysynthetic language, Ojibwe has morphologically complex inflection for verbs and nouns, and productive derivational morphology, leading to hundreds of thousands of unique possible word forms. Our goal is to create an accessible, adaptable, and accurate FST-based model of this morphological complexity in Ojibwe, which will be applied for a variety of purposes in the domains of pedagogy, research, and general use, including a morphologically tagged text-based corpus, a spell-checker, an automatic verb conjugation tool, and an intelligent dictionary search tool. While focused primarily on the creation of the core *OjibweMorph* FST, we devised a modular, language-neutral system for creating an FST from a set of human- (and machine-) readable spreadsheets. We see this as a useful case study for future researchers and language activists who may wish to create FSTs in other languages without the need for technical expertise.

The paper is organized as follows. Section 2 gives a general background on the use of FSTs in the context of Indigenous languages of North America. In Section 3 we detail the key features of the Ojibwe language relative to the current project. Section 4 gives an overview of our modeling approach, and Sections 5–7 detail the three modules that combine to generate the *OjibweMorph* FST: (i) the morphological module, also called `OjibweMorph`, (ii) the lexical module `OjibweLexicon`, and (iii) the language-neutral compilation module `FSTmorph`. Section 8 describes the evaluation results of the FST's performance in terms of coverage and accuracy. Finally, in Section 9 we discuss current and future applications of the *OjibweMorph* FST, and our approach more generally. Section 10 concludes.

2 Background: Finite-state transducers (FSTs)

2.1 A primer

The goal of this section is to give a high-level overview of finite-state transducers (FSTs). To maximize accessibility, we will focus for now on examples from English. FSTs can bidirectionally map between two sets of symbol strings, notated with the pair `upper:lower`, where `upper` is one set of symbols that are mapped to or from the second `lower` set of symbols. They are one of the most common tools for creating a morphological analyzer, which generally defines a mapping between a set of orthographic sequences and a set of grammatical tags. For example, a very simple FST for a snippet of English could *analyze* or *lookup* an inflected word like `walked` as `walk+Verb+Past`. Correspondingly, if given the set of tags `walk+Verb+Past`, an FST can *generate* or *lookdown* the form `walked`. Such an FST might consist of the pairs `walk+Verb:walk` and `+Past:ed`.

One of the most popular, fully open-source toolkits for creating FST-based morphological analyzers is *foma* (Hulden, 2009). *Foma* uses two formalisms originally developed by Xerox to encode lexical information (*lexc* formalism) and phonological rules (*xfst* formalism) (Beesley & Karttunen, 2003). In the lexical component, one can specify sets of stems like `walk`, `work`, and `bake` and a set of affixes like `-ed`. The phonological rule component provides a way to model morphophonological and orthographic alternations. For example, to generate the proper output for a lookdown of `bake+Verb+Past`, we could specify a rule that deletes the `e` in the affix `-ed` when the stem to which it attaches ends in an `e`. This would ensure the correct generation of `baked` rather than the incorrect generation `bakeed` in the absence of such a rule.

A final relevant feature of *xfst*-style FST compilers such as *foma* is the ability to encode long-distance dependencies through the use of *flag diacritics*, which introduces a limited system of memory into the model (Beesley & Karttunen, 2003, ch. 8). Long-distance dependencies are cases where two non-adjacent symbols are linked in some way. For example, when a particular prefix either must (or must not) occur with a particular suffix, with any number of different stems or other morphemes in between. To take a simplified version of the example from Mans Hulden's *foma* tutorial (https://fomafst.github.io/morphtut.html), the English prefix `un` often

must co-occur with the suffix `able` when attached to certain verbs like `walk`. For example, we can say `unwalkable` but not `unwalk`. Abstractly, we can use flag diacritics to indicate that, when the prefix `un` is specified, we must ensure that somewhere down the line the suffix `able` will appear. While it is always possible to create an FST without the use of flag diacritics, they allow for a much more compact representation of the lexicon. Use of flag diacritics comes at the cost of decreased parsing speed, but if used judiciously this tradeoff is not severe enough to dampen the usability of a model.

2.2 FSTs for Indigenous languages

FSTs in general have a number of properties that make them well-suited for creating technologies for Indigenous or minority languages. First, the ability to both analyze and generate with a single model provides a greater range of applications than a model that can only do one or the other. Second, FSTs do not require large datasets for training, as these models are hand-coded by language experts with the aid of existing documentation such as grammars and dictionaries. Third, the deterministic quality of the most basic FSTs ensures that models generate consistent results (though note that probabilistic variants of FSTs are also in common use). As long as the underlying specifications within the FST are accurate, then the outputs of the FST itself will be accurate. This is of special importance for the common case where FSTs are used by language learners. Fourth, FSTs are well established computational tools: There exist several open-source implementations of compilers for creating them, making use of various computational algorithms enabling efficient and fast compilation, and they can be turned into computational packages that can easily be integrated within other applications for providing various linguistic functionalities, e.g. spell-checking, word-form analysis, or inflectional paradigm generation (Arppe, Lachler, Trosterud, Antonsen, & Moshagen, 2016).

There have been a number of successful efforts to build FSTs for Indigenous languages of North America, and our work directly relies on many of the insights and lessons within this growing body of work. This includes models of multiple languages within the Algonquian family including Odawa (Bowers, Arppe, Lachler, Moshagen, & Trosterud, 2017), Plains Cree (Harrigan et al., 2017; Snoek et al., 2014), East Cree (Arppe, Junker, & Torkornoo, 2017), Blackfoot (Kadlec, 2022), Arapaho (Kazeminejad, Cowell, & Hulden, 2017), and the French-Cree mixed language Michif (Davis, Santos, & Souter, 2021), as well as from other North American languages including Tsuut'ina (Arppe, Cox, et al., 2017; Holden, Cox, & Arppe, 2022), Northern Haida (Lachler, Antonsen, Trosterud, Moshagen, & Arppe, 2018), San Mateo Huabe (Tyers & Castro, 2023), and Gitksan (Forbes, Nicolai, & Silfverberg, 2021), to name a few. Outside of North America there are many more examples, such as models for Kunwinjku (Lane & Bird, 2019) and Nen (Muradoğlu, Evans, & Suominen, 2020). Most relevant to the current project is the work of Bowers et al. (2017), as Odawa is classified as an eastern dialect of Ojibwe. We discuss this model in various points throughout the paper, so we will not draw out specific details here.

As the above paragraph should demonstrate, there is a robust body of work applying FST technology to Indigenous languages, including a dialect of Ojibwe. As such, besides detailing yet another example of the fruitfulness of FSTs, the current paper provides a general framework that increases the accessibility of building such models. In a nutshell, anyone familiar with editing spreadsheets and tables should be able to make a model using our tools. However, it should be noted that all languages still have a base complexity that ultimately must be reflected in the model: regularities, sub-regularities, and exceptions must all be woven together. The spreadsheets make the encoding of these different facets of morphophonology straightforward, but there is no getting around the need for significant language-specific expertise to build a full model. Furthermore, the approach lends itself well to modeling variation between dialects in a straightforward way—a necessary feature for any approach that wants to best serve Indigenous language communities. We are therefore swimming in the same direction as other recent work such as Littell, Stewart, Davis, Pine, and Kuhn (2024)'s Gramble, a tabular programming approach to building morphological parsing models. We were unaware of this new work during the development of our system. Our approach is distinct in that the end result builds a typical"two-tape" FST, whereas Gramble uses a "multi-tape" system. It is not yet clear whether the resulting modeling from the Gramble system has the same immediate

affordances as ours, which can be seamlessly integrated into existing systems for spell-checkers and other tools for learning (see Section 9 for details), but we believe the existence of this work affirms the general approach we have taken. In any case, our main goal in this paper is to describe the system relative to Ojibwe, rather than its general affordances across languages.

2.3 Concatenation versus chunking in FSTs

Following the discussion in Bowers et al. (2017, Sec. 5.1), we can classify existing models on a continuum from *concatenative* on one end, where individual affixes are listed in the lexicon, to *chunked* on the other, where internally complex sequences of affixes are listed in the lexicon. To take another example from English, `walkability` is technically composed of the stem `walk` and the underlying suffixes `able` and `ity`, as in the pair `walk+able+ity:walkability`. In a concatenative model, each of these individual suffixes would be specified in the model as `+able:able` and `+ity:ity`, the lexicon would be set up to ensure they are concatenated in the proper order, and some set of phonological rules would be specified in order to reach the correct surface form of `ability` rather than `ableity`. In a chunked model, one would instead specify `+able+ity:ability` as its own form. This obviates the need for specifying concatenation order and phonological rules that apply within the suffix complex—a complex form can simply be listed as is.

Each of these approaches has benefits and drawbacks, and most models fall somewhere in the middle rather than at the extremes. A first consideration is the general type of inflection dominant in a given language. In highly *fusional* languages, where it is difficult to distinguish morpheme boundaries, or in languages where the morphophonology or orthography obscures morphological boundaries in complex ways, chunking becomes a near necessity. In *agglutinating* languages with consistent orthographic conventions, where morpheme boundaries are generally clear, it becomes possible to consider a concatenative approach. Relatedly, the approaches differ in how they deal with irregular or sub-regular forms (see also Fransen, 2020, for an interesting discussion on this topic in the case of stem allomorphy in Old Irish). The chunking approach can easily capture exceptional patterns, as all forms are hard-coded and listed anyway. In a concatenative approach, since forms are derived in a generative fashion by concatenation and phonological rules, exceptions and irregularities need to be carved out by overriding these general rules, creating complexity.

Another difference between the approaches is their generative capacity. Concatenative models have increased generative capacity compared to a chunking approach. As noted by Bowers et al. (2017), this allows for "missing cells" within descriptions of morphological paradigms to be filled in automatically by the model. In a chunked approach, such missing forms will either have to be filled in by hand with an educated guess, or left out of the model entirely. However, automated guessing also opens the model up to generating illicit forms in these gaps, and requires significant knowledge about the morphophonology of a given language, and command of the computational rule formalism. Put differently, building concatenative models requires an advanced level of metalinguistic understanding that is not available for many minority languages, while simple descriptions of paradigm tables are often readily available. If complete paradigms have been created (by hand) for every inflectional type within a given language, this information can be used to generate *lexc* content following the chunking approach, as was done for suffixation in the case of East Cree (Arppe, Junker, & Torkornoo, 2017). This is similar to the approach we will take in this paper.

A final difference is in the "compactness" of the description of forms. The concatenative approach generally leads to a more concise description than chunking. For the purpose of illustration, imagine a language that has five distinct suffix slots, each with four possible forms within a given slot. Assuming all possible combinations are valid, the result is 1,024 (4^5) possible combinations. A chunked model would require listing all 1,000+ possible forms individually, while a concatenating model could do so through a relatively small set of lexical forms and rules. In an ideal world, concatenative models should therefore be easy to change and maintain. For example, if a change is needed to the form of a specific morpheme to capture some sort of variation in a dialect or spelling convention, in principle this can be accomplished by making a change to a single location in the lexicon, which will then be automatically dispersed to all places where that morpheme occurs in the language. In a chunking model,

such a change would require touching the many individual forms hard-coded into the model. However, in practice, making changes in a concatenative model can get complicated.

Indeed, the project presented in this paper started with the aim of adapting Bowers et al. (2017)'s concatenative model of Odawa for other dialects of Ojibwe. However, the complexity of the model made this task difficult, leading to the approach detailed in the present paper. Modeling the full morphophonology of a language requires dozens of rules, many of which interact in opaque ways. Changing a single form or rule can have unintended knock-on effects that are difficult to diagnose or constrain. Similarly, in the effort to create a maximally concise description, it is often necessary to increase the level of abstraction in forms and rules. This can decrease human readability and comprehensibility, especially for non-experts. We will have more to say about this when we introduce our own approach to modeling in Section 4, which is largely a chunking-based approach.

3 Background: Ojibwe language

This section covers some basic language background necessary for understanding both the general context of the work as well as the specific design choices that were made in the course of creating the FST and associated framework. We first detail the geographic distribution and vitality of the various dialects of Ojibwe. We then describe the phonological and orthographic systems. Finally, we cover the basic properties of the complex morphosyntax of Ojibwe.

3.1 Vitality

Ojibwe is not a single language, but rather a cover term for a set of closely related and largely mutually intelligible dialects spoken in the land area around the Great Lakes of North America, from southwestern Quebec westward to outlying communities in Alberta and British Columbia within Canada, and from Michigan westward to outlying communities in North Dakota and Montana within the US. Commensurate with the population of Ojibwe people, most Ojibwe speakers today are located in what is now the province of Ontario, but there are sizable communities of speakers and learners all across Ojibwe country. Current estimates from the Statistics Canada 2021 census cite the number of Ojibwe speakers at 25,440, with likely not more than a few thousand additional speakers being located in the US (the most recent estimate from the American Community Survey from 2006–2010 puts the number at 8,371). While the language is at risk due to the impacts of colonization, especially the ongoing effects from residential school systems, there are many efforts to ensure the vitality of Ojibwe into the future. These include child and adult language immersion programs, language nests, and dictionaries and books in digital and printed formats, as well has various technologies including text-to-speech systems (Chan & Hammerly, 2025; Hammerly et al., 2023; Wang et al., 2025) and a machine translation model (Nguyen, Hammerly, & Slifverberg, 2025).

3.2 Dialects and variation

Being spoken over a large geographic area, with hundreds of distinct First Nations and Tribal Nations in multiple provinces and states, there is variation across Ojibwe varieties that ranges from the macro to the micro, including lexical, phonological, and grammatical differences (Valentine, 1994, 2001). One goal of our project is to build a system that is capable of being flexible enough to deal with micro-variation within the most closely related dialects, as well as have the flexibility to be applied to more divergent dialects and other languages in the Algonquian family and beyond. The goal of this section is to set out a specific dialect group that we aim to capture, to narrow the scope of our project.

The most obvious split within Ojibwe dialects is in the degree of *vowel syncopation*, a phonological process that arose starting in the early part of the 20th century whereby short unstressed vowels are reduced or deleted (Bloomfield, 1957, p. 5). Syncopation is most common in Eastern dialects such as Odawa, and has a significant impact on the surface form of many words.

In the current work we focus on non-syncopating dialects. Specifically, we aim to model the group of closely related varieties commonly known as Southwestern Ojibwe (ISO 639-3:

ciw), which we take to include those variants of Ojibwe spoken in Minnesota, Wisconsin, and Northwestern Ontario along the Rainy River.[1] The remainder of this section, and much of the rest of the paper, therefore focuses on describing and modeling the Southwestern Ojibwe group, and in the interest of brevity we will use the term "Ojibwe", unless disambiguation is needed, to refer to this specific sub-variety. We return to some of the promises and challenges of expanding or adapting the current model to capture a wider variety of dialects in Section 9.

3.3 Orthography and phonology

Ojibwe is traditionally a purely oral language, with widespread writing systems emerging from missionaries in the context of colonial contact over the course of the 19th and 20th centuries. Today, there are two main types of writing systems in use—an alphabet of Latin origin and a syllabary based on the Canadian Aboriginal Syllabics—with significant variation within each of these categories. Southwestern Ojibwe is almost exclusively written in a Latin-based variant known as the Double Vowel system, first devised by Charles Fiero in the 1970s, and popularized through the documentary work of John Nichols in the following decades. Given this widespread adoption within our target dialect group, our FST is designed with these conventions in mind, and the remainder of this section is focused on detailing this system and its broad relation to the Ojibwe phoneme inventory.

The inventories of vowels and consonants are summarized in Table 1. All vowels except /e/ come in long/short pairs, and all long vowels have nasal counterparts. Nasal vowels are of particularly uncertain phonological status as discussed briefly in Nichols (1980, p. 6-7). In any case, it is represented in the orthography and so deserves special attention. Stops, fricatives, and affricates come in voiced/voiceless pairs, with the voiced counterparts commonly referred to as the *lenis* consonants and the voiceless the *fortis* consonants (e.g. Swierzbin, 2003).

Short Oral	Long Oral	Long Nasal
a	aː	ãː
	eː	ẽː
i	iː	ĩː
o	oː	õː

	bilabial	dental	alveopalatal	palatal	velar	glottal
stop	p b	t d			k g	ʔ
fricative		s z	ʃ ʒ			
affricate			ʧ ʤ			
nasal	m	n				
glide				j	w	

Table 1 Broad transcription of vowel and consonant inventory of Ojibwe. Based on Swierzbin (2003).

The orthographic symbols used for vowels and consonants are given in Table 2. Their correspondence with the phonemes is mostly intuitive for those already familiar with the Latin alphabet, with the exception of the glottal stop, which is represented with an apostrophe (apostrophes are not otherwise used as punctuation in the writing system). Also important to highlight is that *e* is a long vowel despite being represented with a single character. As there is no short variant of this vowel in Ojibwe, this does not give rise to any ambiguities. In general, the writing system is used as a broad transcription system. Rather than standardized spellings for specific words across dialects, the writing system generally reflects the way a particular dialect is actually spoken, with sound-symbol correspondences being consistent across communities. For example, the spelling of the word for "one month" varies between *ingo-giizis* and *ningo-giizis* along with the speaker's pronunciation.[2] This presents a technical challenge

[1] Dialects spoken along the Rainy River are commonly known as Border Lakes Ojibwe. This group has no specific ISO classification, but for current purposes it can be classed with the other ciw varieties. As with all varieties within a dialect, there are a number of known markers that diverge from the rest of the dialect group. However, the grouping can be motivated by geographic proximity, long-standing community connections, as well as linguistic similarities. Furthermore, the Border Lakes Variety is represented along with dialects from Minnesota and Wisconsin within the Ojibwe People's Dictionary, which as we detail below is a primary lexical source for our FST. This demonstrates the utility of treating these varieties as a unit.

[2] See OPD entries: https://ojibwe.lib.umn.edu/main-entry/ingo-giizis-adv-num and https://ojibwe.lib.umn.edu/main-entry/ningo-giizis-adv-num.

Orthography	Phoneme
a	/a/
aa	/aː/
aanh	/ãː/
i	/ɪ/
ii	/iː/
iinh	/ĩː/
o	/o/
oo	/oː/
oonh	/õː/
e	/eː/
enh	/ẽː/
p	/p/
b	/b/
t	/t/
d	/d/
k	/k/
g	/g/
’	/ʔ/
s	/s/
z	/z/
sh	/ʃ/
zh	/ʒ/
ch	/ʧ/
j	/ʤ/
m	/m/
n	/n/
w	/w/
y	/j/

Table 2 Correspondences between orthographic symbols and phonemes in the double vowel writing system.

for our project, since we need to ensure enough flexibility to be inclusive across a range of different pronunciations/spellings.

There are two additional orthographical conventions that fall outside of representing the phonemic inventory or otherwise deserve a bit more explanation (see also Sullivan, 2016, p. 83). First, there is the use of “h” in the spelling of certain exclamatory particles such as *howah* ‘wow!’, *ahaw* ‘ok’, and *hay* ‘darn’. The orthographic segment *h*, generally pronounced [h], does not occur on its own outside of this limited class of words (though occurs as part of a multicharacter symbol with *sh* and *zh*, for example).

Second, the combination of *nh* occurring after a long vowel is used to encode word-final nasal long vowels lexically present in words such as *abinoojiinh* ‘child’, pronounced [abɪnuːʤĩː]. When suffixes are added to these words, for example the animate plural suffix allomorph *-yag* as in *abinoojiinyag* ‘children’, pronounced [abɪnuːʤĩːjag], spelling conventions generally dictate the deletion of the *h*.[3]

3.4 Morphology

As a polysynthetic, largely agglutinating language, Ojibwe has a complex system of morphosyntax and morphophonology. We detail the basic properties of the system here, with more detail to come during the presentation of the morphological module of our FST in Section 5. We split our discussion into two parts: nominal morphology and verbal morphology.

3.4.1 Nominal morphology

The first important fact about Ojibwe nouns is that the language has a robust, bipartite noun class system based roughly on the semantic animacy of a noun. All lexical nouns fall into

[3] In actuality, this description is slightly misleading. The leading analysis of words ending in a nasalized long vowel is that the stems end in the glide [j], orthographically represented as *y*. A general phonological rule that deletes glides word-finally results in the singular stem ending in just a nasal vowel on the surface. Adding the plural suffix *-ag* bleeds this rule, allowing the glide from the stem to surface.

either the ANIMATE (NA) or INANIMATE (NI) class, with a number of nominal stems showing alternations between the two classes as well (see the example (1) below). As indicated above, a common notational shorthand within the descriptive literature on Algonquian languages for animate nouns is NA (= noun animate) and for inanimate nouns is NI (= noun inanimate). In modern Ojibwe there is no reliable marking of animacy on bare noun stems, but certain nouns like *makwa* 'bear (NA)' and *mishi* 'firewood (NI)' have retained the animate singular marker *-a* and inanimate singular marker *-i* for phonological reasons.

The number system of Ojibwe distinguishes between SINGULAR (SG) and PLURAL (PL), and the number suffix differs depending on the animacy of the noun. Along with other elements such as demonstratives and verb agreement, number is a clear indicator of animacy. A minimal pair showing this distinction with the same basic root is given in (1), where the animate variant translates to "trees" and uses the plural suffix *-oog* and the inanimate variant to "sticks" and uses the plural marker *-oon*.

(1) a. mitig-oog
wood-ANIM.PL
'trees (NA)'

b. mitig-oon
wood-INAN.PL
'sticks (NI)'

Plural marking in particular is useful in revealing the patterns of regular allomorphy between the stem and nominal affixes, which generally depend on the phonological properties of the right edge of the stem. These types of alternations have lead to the categorization of stems into different PARADIGM CLASSES (Nichols, 1980; Valentine, 2001). As will be discussed in more detail in Section 5, paradigm classes play a key role in the organization of the inflectional module of our FST generation architecture, and so deserve a bit of attention here. In total, there are 12 nominal paradigm classes. We refer the reader to Steiner and Hammerly (to appear) for a recent overview of the role of these classes in Ojibwe and only give a sense of the system here.

Consider the examples in (2) and (3), where the plural marker in (2-b) appears as *-oog* and the marker in (3-b) surfaces as *-ag*. Additionally, in (3-b) we see the surfacing of a *w* on the stem that is not present in the singular form in (3-a), resulting in the string *ikwewag* for the plural of *ikwe*.

(2) a. mitig
wood.ANIM.SG
'tree (NA)'

b. mitig-oog
wood-ANIM.PL
'trees (NA)'

(3) a. ikwe
woman.ANIM.SG
'woman (NA)'

b. ikwew-ag
woman-ANIM.PL
'women (NA)'

The stem in the examples in (2) belongs to class 4a, which all end in a consonant followed by a *w*. That is, the *underlying* form of the stem, as represented by the orthography, is */mitigw/*. The stem in (3) belongs to class 2, which all end in a long vowel followed by a *w*. That makes the underlying form of this noun */ikwew/*. In both cases, in the singular, the glide is subject to a general word-final deletion rule, so does not surface. In turn, plural marking is thought to have the underlying form *-ag*. Stems in class 4a, as in (2-b), are subject to a contraction rule that turns the sequence *wa* into *oo* specifically when it follows a consonant (other than "k"). This leads to the appearance of the *-oog* allomorph of the animate plural. With the stems in (3-b) this rule does not apply since the *w* is preceded by a vowel, not a consonant, leading to the "regular" form of the plural marker *-ag*.

Returning to the types of inflections that nouns can take, within animate nouns, there is another important set of markers that indicate a function known as OBVIATION, which is fused with the number marking system. Obviation has a complex distribution, but at a basic level, it expresses the discourse or syntactic prominence of animate nouns. Nouns that are PROXIMATE (unmarked in the singular, *-ag* in the plural) are generally more discourse prominent. The proximate form is also used as the default in cases where there is only one third person noun in the discourse. Nouns that are OBVIATIVE are generally less discourse prominent, and in

the singular marked with a suffix *-an* (and its allomorphs). Many dialects of Ojibwe do not show a distinction in number with obviative nouns, so use *-an* for both singular and plural. However, the main variety targeted for the current project (Border Lakes Ojibwe) has retained the contrast in number, and uses the marker *-a'* for obviative plural nouns. Inanimate nouns in Ojibwe do not show any contrasts for obviation.

Besides animacy, another broad split in the categorization of nouns is between DEPENDENT (NAD/NID) and INDEPENDENT (NA/NI) nouns. Dependent nouns are obligatorily possessed, and therefore always are inflected for some degree of possessive morphology and never appear in their bare stem form. As indicated above, this is generally notated with NAD (= noun animate dependent) and NID (= noun inanimate dependent). Dependent nouns are generally those involving family relations, body parts, and important animals or objects. Possession, for both dependent and independent nouns, is most reliably marked by the presence of a PERSON PREFIX and PERSON/NUMBER SUFFIX that indicates the person, number, and obviation of the possessor, as shown with a second person plural possessor in (4).

(4) *gijiimaaniwaa*
gi-jiimaan-iwaa
2-canoe.NI-PL
"your (PL) canoe"

Many independent nouns additionally take a POSSESSIVE SUFFIX (POSS) *-im*, as in (5). This suffix is obligatory on certain independent possessed nouns, but is generally reported to be ungrammatical on all dependent nouns. The distribution of this marker is not currently well understood.

(5) *gizhiishiibim*
gi-zhiishiib-im
2-duck-POSS
'your duck'

There are a number of other inflectional affixes that can be marked on nouns. One of the most common is the DIMINUTIVE (DIM) marker, most typically realized as *-ens*, which generally indicates the small size relative to its kind, but is also commonly lexicalized to some less compositionally transparent meaning (e.g. the diminutive of *ishkode* "fire" is *ishkodens*, meaning "match" rather than "small fire"). The PEJORATIVE (PEJ) *-ish* (6) indicates that something or someone is disfavored, but it can also be used affectionately, especially when combined with the diminutive. The PRETERIT (PRT) *-iban* (7) indicates that the referent is deceased. It can also, somewhat archaically, be used on inanimate items to indicate it is broken or lost. The LOCATIVE (LOC) *-ing* (8) indicates spatial relationships or similarity. Finally, the VOCATIVE (VOC) *-dog* (9) is a plural address form. It is generally used on animate nouns referring to humans, but can also be used on inanimate nouns.

(6) *jiimaanish*
jiimaan-ish
canoe-PEJ
'that ol' canoe'

(7) *nimaamaayiban*
ni-maamaay-iban
1-mother-PRT
'my late mother'

(8) *waakaa'iganing*
waakaa'igan-ing
house-LOC
'in, at, to, like the house'

(9) *Anishinaabedog!*
anishinaabe-dog
Ojibwe-VOC
'My fellow Ojibwe people!'

It is also possible for multiple pieces of nominal inflection to stack together. These stackings occur in a set order, can condition sound changes between suffixes, and are subject to certain restrictions. The template for the nominal morphology is summarized in Table 3. We will not review the full set of restrictions here for reasons of space, but as an example: it is not possible to have both the locative and the number/animacy/obviation markers at the same time—in other words, the two morphemes compete for expression within the same slot. This slot is

PERS. PREFIX	STEM	DIMINUTIVE	POSS.	PEJORATIVE	PERS. SUFFIX	PRETERIT	BASIC

Table 3 Template for nominal morphology. The basic suffix is the locus of expression for the locative, vocative, and number/animacy/obviation markers, which are in complementary distribution and therefore never appear at the same time.

known as the BASIC suffix, and is also the locus of expression for the vocative marker (which is therefore also in complementary distribution with the locative and number/animacy/obviation markers). These restrictions are part of what we must model within the FST to ensure only well-formed sets of inflection can be generated.

3.4.2 Verbal morphology

The intricate system of verbal morphology in Ojibwe has been a major focus of documentation and study, and presents one of the biggest challenges for second language learners of the language. Modern descriptions of the system largely stem from the work of Leonard Bloomfield (e.g. Bloomfield, 1957), who refined the paradigmatic classification of the verbal system that continues to be in active use today.

The most fundamental split in the verbal system is something that we will refer to as PARADIGM CLASSES, which divides verb stems and their corresponding sets of inflections based on transitivity and grammatical animacy. There are four basic paradigm classes (note in each case the "V" stands for "Verb"): VII (Inanimate Intransitive), VAI (Animate Intransitive), VTI (Transitive Inanimate), and VTA (Transitive Animate). With intransitives, animacy is relative to the sole argument of the clause. With transitives, animacy is relative to the object. For example, in a VTA, the object is always grammatically animate, but the subject may be either animate or inanimate.

In addition to the four basic paradigm classes, there are a number of finer-grained divisions that must be made. For example, certain VAI and VII stems that *only* allow for plural arguments because they inherently refer to groups or collections of things (e.g. *okoshinoog* 'they (animate) lie in a pile', where a pile implies there is a plurality of things) belong to their own class (VAIPL and VIIPL, respectively). Perhaps the most common additional class is the VAIO (Animate Intransitive + Object) paradigm. These are stems that are *derivationally* formed as VAIs, so appear to be intransitive (hence, the VAI classification). However, they allow for transitive *inflection* that overlaps with the set used for VTAs and VTIs, and they can take overt nominal objects (hence the "O"). Unlike regular transitive verbs, these stems only allow for third person objects (either animate or inanimate). Furthermore, like the nouns, verbs in Ojibwe are split into different inflectional classes based on the phonological properties of the right edge of the stem.

Another important split in the classification of verbal morphology is ORDER, which defines distinct paradigms of verbal inflection roughly corresponding to different clause-types. Ojibwe has three basic inflectional orders: the INDEPENDENT (IND), which generally occurs as a main clause, the CONJUNCT (CNJ), which generally occurs in embedded clauses and most questions, and the IMPERATIVE (IMP), which is the command form of the verb. In the current model, we also conceptualize the PARTICIPLE (PCP) forms as a distinct inflectional order, though the morphology overlaps significantly with the conjunct order. Participle forms are generally used in certain Southwestern varieties to form relative clauses (Sullivan, 2016). For current purposes, the syntactic distribution of the these different orders is less important than detailing the significant inflectional differences between each one.

The independent order hosts up to four basic inflectional affixes descriptively known as the PERSON PREFIX, THEME SIGN, CENTRAL AGREEMENT, and PERIPHERAL AGREEMENT. Without going too far into the complexities (while still giving a good taste of it), these four slots combine to indicate the animacy, person, number, and obviation of the verbal arguments. For example, the verb form in (10) has the first person prefix *ni-* indicating the presence of a first person in the clause, the general third person theme sign *-aa* indicating that there is a third person object (and also, by elimination, making it clear that the first person is the subject), the first person plural central agreement *-naan* that specifies the number of (in this case) the subject, and the third person proximate plural peripheral suffix *-ig* that indicates the number and obviation of (again, in this case) the third person object.

(10) *niwaabamaanaanig*
ni-waabam-aa-naan-ig
1-see-3-1PL-3PL
'we (EXCL) see them (PROX)'

The conjunct order has only suffixal basic inflectional morphology, with the theme sign (in transitive clauses) and the central agreement appearing. Neither the person prefix nor peripheral suffix are ever realized, but the central agreement marker often consists of several discernible morphemes. Consider the example in (11), which has the same basic argument combination as (10) (but is in the negative, since it shows certain patterns more clearly). We see the same theme sign *-aa* indicating there is a third person object. The central agreement can be split in two: *-ang* indicates there is a first person exclusive argument, and *-idwaa* indicates the third person is proximate and plural.

(11) *waabamaasiwangidwaa*
waabam-aa-siw-ang-idwaa
see-3-NEG-EXCL-3PL
'if/when we (EXCL) do not see them (PROX)'

Like the conjunct order, the imperative order only contains suffixal inflection for arguments. Imperative forms only allow for the subject to be a second person, so the argument combinations are more restricted and there are no imperative forms for the VII paradigm classes. An example of a command with a VAI verb is given in (12). The *-n* suffix indicates that the command is oriented towards a singular second person.

(12) *Biindigen!*
biindige-n
enter-2SG
"Come in!"

Finally, the participle form of the verb is used is certain dialects when forming a relative clause. There is one important grammatical concept needed as background to understand the form of participles: a vowel ablaut process known as INITIAL CHANGE that targets the first vowel of the verb stem. Initial change can occur on conjunct order verbs (not just participles) to form what is known as the CHANGED CONJUNCT. The changed conjunct appears primarily when *wh*-questions have occurred. This can be seen in (13) (taken from the OPD), where the verb *boopoogidi* "s/he farts off and on" has undergone initial change. This is evident in the fact that the first vowel is realized as "waa" rather than "oo". In terms of inflection, there is no difference between the unchanged (often called *plain*) and changed conjunct.

(13) *Awenen bwaapoogidid?*
awenen IC.boopoogidi-d
who IC.fart-3SG
"Who keeps farting?"

There is a lot of overlap between the changed conjunct and participle forms. In many cases, the two are string identical, but in certain parts of the paradigm they pull apart (for a comprehensive review, see Sullivan, 2016). For illustrative purposes we will focus on a case where they are distinct. Like the changed conjunct, all participles involve initial change. What makes them distinct is a special agreement marker that indicates the number and obviation of the head of the relative clause. Consider the pair in (14), which both have an animate third person plural proximate subject. The form in (14-a) is the changed conjunct, where the plural is marked by the central agreement *-waa*, while the form in (14-b) is the participle, where plural is marked by the peripheral suffix *-ig* (which also triggers palatalization of the "d" to "j").

(14) a. *baandigewaad*
IC.biindige-waa-d
IC.enter-3PL-3
"after they entered"

b. *baandigejig*
IC.biindige-d-ig
IC.enter-3-3PL
"the ones who enter"

PERS. PREFIX	STEM	THEME SIGN	NEGATIVE	CENTRAL AGR.	MODE	PERIPHERAL AGR.

Table 4 Template for verbal morphology. The independent order can show the full set of markers, while the conjunct order lacks the person prefix and peripheral agreement. This template does not apply straightforwardly to the imperative order.

It is important to note that not all Southwestern varieties use participle forms. For example, in Border Lakes Ojibwe, relative clauses are instead formed by using a relativizing preverb *gaa-* paired with a regular conjunct order verb, as in (15).

(15) *gaa-biindigewaad*
gaa-biindige-waa-d
REL-enter-3PL-3
"the ones who enter"

In addition to inflection that marks the animacy, person, number, and obviation of the verbal arguments, verbs can also be marked for suffixes related to NEGATION (NEG), which indicates whether the polarity of the predicate is positive or negative, and MODE, which conveys information related to tense, aspect, and modality. For the independent and conjunct orders, there are four possible modes (definitions paraphrased from Valentine, 2001): (i) NEUTRAL (NEU), which indicates a straightforward assertion of truth; (ii) PRETERIT (PRT), which indicates an event took place in the past and is no longer ongoing; (iii) DUBITATIVE (DUB), which indicates doubt or uncertainty regarding the assertion; and (iv) the PRETERIT-DUBITATIVE, which generally marks an event as occurring in the past where the speaker does not have first-hand knowledge. The imperative order has a different set of modes: (i) SIMPLE (SIM), indicating a command should be carried out in the present moment; (ii) DELAYED (DEL), indicating the command should be carried out at a later time; and (iii) PROHIBITIVE (PRB), indicating a negative command (e.g. don't run!).

By way of summary, the maximal inflectional template for Ojibwe verbs is given in Table 4.

4 Overview and approach

The goal of this section is to give detail on the team that has worked on this project, report the methods for gathering and verifying data to support the creation of the FST, including the process of seeking permissions for the use and sharing of certain data, and give an overview of the guiding principles and architectural design of the tools we have created.

4.1 Team and collaborators

The core team creating this FST consisted of a documentary linguist specializing in Ojibwe lexicography who serves as the editor of a major online dictionary, a theoretical linguist and fieldworker who is a member of the White Earth Nation in Minnesota and specializes in morphology and syntax, and three computational linguists with a variety of specialties and deep experience in creating computational tools for Indigenous and low-resource languages.

In addition, a highly fluent and respected elder from Nigigoonsiminikaaning First Nation in Northwestern Ontario worked with the fieldworker on our team to ensure the accuracy of the target morphological forms for generation. Other members of the community including language instructors have also been consulted on both a formal and informal basis throughout every stage of the project from conception to its current state of development. As discussed further in Section 9, one of the major drivers of development and design was to incorporate the FST into the Ojibwe People's Dictionary to create intelligent dictionary search, where users can input morphologically complex words and return a result that gives the relevant entries in the dictionary and an analysis of the original word.

The team has also consulted at various points with collaborators in the educational technology sector working on a language revitalization platform called *Anishinaabemodaa! Waking up Ojibwe*. As we detail in Section 9, one of the ways the FST will be used is as an online automatic verb conjugation tool. It was important for us to make design choices that would ensure these sorts of applications and extensions of the FST are readily possible.

4.2 Data sources

The data for the creation of the inflectional module of our system comes primarily from fieldwork conducted between 2017 to the present in Northwestern Ontario and Minnesota, but also makes use of published grammatical descriptions and unpublished work related to the documentation of the language shared with the authors, most notably Valentine's (2001) grammar of the Eastern Ojibwe dialect Nishnaabemwin, Valentine's (2024) web-based description of Ojibwe, Nichols' (1980) dissertation describing the morphology of Mille Lacs Ojibwe, and Nichols' unpublished grammar of Southwestern Ojibwe (cited as Nichols (2011)). These prior grammatical descriptions were used as one starting point for primary fieldwork, where the patterns and generalizations were independently confirmed. We presented a form based on what is known from existing sources, or in some cases our personal knowledge of the language, and asked (i) if the speaker recognized the form, and (ii) what the translation of the form is. In cases where the speaker did not recognize the form based on previous descriptions or our educated guess, we reverted to translation to fill in the cell of the paradigm (e.g. asking questions like "how do you say 'I see you' in Ojibwe?"), and re-confirmed the form again in a later session to be sure of its shape and meaning.

The lexical module is primarily comprised of data from the Ojibwe People's Dictionary (OPD), an online dictionary of Southwestern Ojibwe (including Border Lakes Ojibwe) created by Dr. John Nichols and currently edited by Nora Livesay (Nichols, 2012). See Section 6.2 for details on the specific data we are making use of. At present, it is hosted by the University of Minnesota Libraries, and its data is publicly available for non-commercial use under a Creative Commons license (CC-BY-NC-SA). The dictionary includes an explicit statement that "Our [the OPD's] goal is to make the dictionary content available as a tool for Ojibwe language revitalization, academic scholarship and cultural awareness." Our use of the data to create the present FST for research and revitalization falls squarely within this mandate. We believe that the relative flexibility with which this data can be used for the purposes of revitalization and scholarship is part of what ensures the present and future vitality of the Ojibwe language, as it allows for the creation of tools such as the one we are presenting here. The lexical module also contains some data from original fieldwork and other published sources such as the glossaries of storybooks. These include place names, proper names, and other items that are not found in the dictionary.

Data sourced from the OPD was also used to test the performance of the model. As detailed further in Section 8, the OPD includes tens of thousands of grammatically tagged inflectional forms, which we used as a test set to ensure the model was producing proper forms and analyses, as well as sentence examples that form the basis of our corpus-based coverage tests.

4.3 Architectural overview

There are three basic modules that comprise our FST generation toolkit: A morphological module (`OjibweMorph`), a lexical module (`OjibweLexicon`), and a compilation module (`FSTmorph`). A schematization of the architecture and compilation flow for the FST can be found in Figure 1. `OjibweMorph` houses morphological paradigms in a spreadsheet format, skeleton *lexc* code, and *xfst* phonological rewrite rules. `OjibweLexicon` houses a lexical database of (at the time of writing) approximately 25,000 words across all parts of speech, mostly sourced, as detailed above, from the Ojibwe People's Dictionary (Nichols, 2012). We have also taken many of our tagging conventions from those established in the OPD, which are in turn in common use among Algonquian language scholars and learners. These two modules are both licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, which allows for others to use, alter, transform, and build upon the materials as long as the result is used for non-commercial purposes, includes attribution to the original source, and is released under the same license. In line with the stated goals of the OPD mentioned above, we hope that this allows the data and framework to be used for educational and research purposes, while protecting it from extractive use that may fall outside of these immediate purposes. We do not aim to be gatekeepers of the language, but we want to honor and respect the contributions and wishes of the speakers and other community members who we have worked with or consulted.

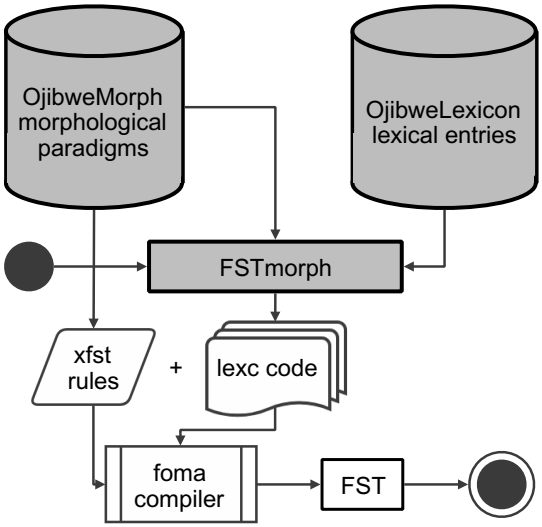

Fig. 1 Compilation of the FST at a glance. Morphological paradigms are described in the repository `OjibweMorph`. The repository `OjibweLexicon` contains lexemes sourced from the Ojibwe People's Dictionary. The compilation code in the `FSTmorph` repository generates *lexc* and *xfst* code from the contents of `OjibweMorph` and `OjibweLexicon` which is then compiled into an FST using the *foma* finite-state compiler.

Our compilation module, `FSTmorph`, generates full *lexc* code from the morphological and lexical databases. These *lexc* files, along with the *xfst* phonological rules, can then be fed to a *foma* compiler, or any other FST compiler implementing the Xerox-style formalisms, e.g. *hfst* (Lindén, Axelson, Hardwick, Pirinen, & Silfverberg, 2011) to generate the final FST. We have released FSTmorph as an easily installable Python package. Since `FSTmorph` does not contain any language-specific data, it is released under a Creative Commons Attribution 4.0 International License, which allows for the materials to be shared and adapted, even for commercial purposes, as long as attribution to the source is given.

The main innovation of the toolkit is the creation of an adaptable framework for *generating* FSTs from spreadsheets and other human-readable and editable formats. We set three primary goals in designing our toolkit: (i) **Accessibility:** ensure that users without prior expertise in the creation of FSTs or other computational tools can create and edit spreadsheets and compile the FST, (ii) **Modularity:** create a system where different components can be independently built and licensed as needed, and (iii) **Adaptability:** create a framework and tools that are adaptable to work with any natural language. While the main focus and scope of the present project and paper was generating a system for Ojibwe, our team is working to expand the approach to other languages and we envision the system as one that researchers outside of our team will be able to make use of with relative ease.

5 OjibweMorph: Morphological Module

With the high-level goals of accessibility, modularity, and adaptability in mind, the goal of this section is to detail the general properties of the morphological model, as well as the specific instantiation for the creation of our Ojibwe FST. This includes the format of our spreadsheets for inflection forms (5.1), prenouns and preverbs (5.2), derivational morphology (5.3), and clitics (5.4).

5.1 Inflectional spreadsheets and phonological rules

As mentioned above, the `OjibweMorph` repository specifies morphological paradigms for the target language along with phonological rewrite rules in the *xfst* formalism which are used to realize morphophonological alternations occurring at morpheme-boundaries. In order to maintain maximal accessibility even for practitioners and community members lacking extensive technical skills, morphological paradigms were implemented as *CSV* spreadsheets that can be

manipulated with commonly available software such as Microsoft Excel. We present the general properties of this spreadsheet structure and rules and the specific instantiation for Ojibwe nouns and verbs.

5.1.1 Basic structure

Our inflectional paradigm spreadsheets list "abstract" inflectional forms, related linguistic features, and a example surface form for all verb types in the language. There is flexibility in that there is no restriction on how many columns can be added, with a configuration file controlling how the custom columns map into the FST (see Section 7 for a high-level overview of this mapping, and the repository readme for technical details). We take advantage of this by specifying different sets of columns for nouns versus verbs to capture the particular types of inflection in each case. We present these additional columns related to nouns and verbs, respectively, in Sections 5.1.2 and 5.1.3. There are, however, a number of columns that are core to the approach and shared across all of the spreadsheets regardless of whether they are related to nouns or verbs:

- **Paradigm**: High-level split based on animacy, as well as transitivity (for verbs) and dependence (for nouns). Modified from OPD POS tag.
- **Class**: Phonological stem classes based on the sounds at the right edge of the stem. Modified from OPD POS tag.
- **Lemma**: Dictionary citation form of the example noun/verb. Modified from OPD citation form.
- **Stem**: Underlying form of the example noun/verb. Modified from OPD stem.
- **Form#Surface**: Surface form of the example noun/verb after all phonological processes have been executed.
- **Form#Split**: Underlying form of the inflection (prior to phonological rules).
- **Form#Source** Source of the surface form.

There can be multiple trios of FORM columns within a given spreadsheet, so the "#" is replaced with a number to indicate correspondences. This allows us to capture certain common patterns of variation within the dialect group. For example, as mentioned in Section 3.4.1, some dialects have a specific form for the obviative plural *-a'* on nouns and verbs (e.g. *zhiishiiba'* means "ducks (OBV)"), while other dialects use the same form *-an* for both singular and plural (e.g. *zhiishiiban* means "duck(s) (OBV)"). We can capture this variation by specifying both possible forms in the row related to the plural obviative in the noun spreadsheets.

One key component of our inflectional paradigm spreadsheets is the SPLIT column which gives a segmentation of the form into *prefix≪stem≫suffix*, for example *gi≪nibaa≫m*, which is the independent, neutral, positive, plural second person subject form of the verb *nibaa* (the resulting form, *ginibaam*, translates to "you all are sleeping"). Note that the prefix or suffix can be empty as necessary. This segmentation is used to factor the form into prefix, stem and suffix sub-lexica in the *lexc* files. The key simplification of this approach is that it allows for the complex morphology in the suffixes to be flattened into a single suffix chunk. Taking verbs as an example, rather than creating a complex set of continuation lexica for each of the five morphological slots (as outlined in Table 4), we can have a single suffix chunk. This means we do not need to model the calculus of concatenation (which includes difficult-to-model phenomena such as metathesis), nor specify any phonological re-write rules related to allomorphy between suffixal elements. All of this can be hard-coded into the model and easily edited if a change is needed.

Each SPLIT form also has a corresponding SURFACE form, which is not used in the model itself, but rather is used to test the output of the model. The SURFACE form shows how the inflection should ultimately appear on the example stem, indicated by the STEM column, that we choose to represent that class. This allows us to validate our model with a set of target forms. This testing is described in detail in Section 8.1.

Each spreadsheet entry is accompanied by an inflection CLASS. The class identifier is used to ensure that stems are combined with prefixes and suffixes of the correct form. With this structure, we can create separate sets of paradigms for each stem class, with the goal of minimizing reliance of abstract forms and phonological rules to increase human readability and

the likelihood that those without linguistic training can understand and edit the paradigms. Recalling our examples in Section 3.4.1, NA stems that end in a consonant followed by a "w" (NA_Cw) show the plural allomorph *-oog* as in *mitigoog* "trees", while NA stems ending with a long vowel followed by a "w" (NA_VVw) show the allomorph *-ag* as in *ikwewag* "women". Rather than creating a single abstract underlying form of the plural morpheme and writing a rule to transform the underlying form to the correct surface form, we have hard-coded this allomorphy into the spreadsheets.

As we will show in more detail as we go through the Ojibwe-specific details, there is still an important role for general phonological rules. Specifically, there are many cases where adding inflection triggers a change to the stem itself. Since we wanted to maintain a single underlying form of the stem for a given word, we have a set of phonological rules that can derive these sorts of alternations to get the correct surface forms. The current version of the model uses a total of 61 phonological re-write rules.

The syntax of *xfst* rules are essentially the same as typical phonological re-write rules, which define what changes, what it changes to, and in what context the change applies (for a primer, see Beesley & Karttunen, 2003, section 1.7). Let us exemplify one such rule as an illustration. With VII verbs ending in "d" (e.g. *zanagad* "it is difficult"), the "d" at the right edge of the stem is deleted whenever the inflection starts with a consonant. Consider the `dDeletion` rule in (16), taken directly from our model.

(16) `define dDeletion d -> 0 || _ SUFBD Cons ;` where...

a. `SUFBD` is a stand-in for the string ≫ that marks the right edge of the stem.
b. `Cons` is the set of consonants in the orthography.
c. `||` can be read as "in the context of".

In other words, this can be read as, "change 'd' to nothing when there is a suffix boundary followed by a consonant to the right".

We also make use of a number of abstract underlying symbols to circumscribe or trigger some of the rules. For example, there is a special type of "n", encoded in the Ojibwe People's Dictionary as "N" within the relevant stems, which palatalizes to "zh" before certain vowels. This is commonly known as the "changeable n", and needs to be distinguished from other types of "n" that do not undergo this palatalization process. We encode this special type of "n" as "n1" within our model to ensure the rule is triggered in all and only the correct places. At the time of writing, we make use of 10 of these symbols, which are fully described within the documentation of the `OjibweMorph` GitHub repository.

Finally, it is also possible to create dedicated spreadsheets for *irregular* forms—forms that cannot be easily modeled through the use of general forms and rules. For example, phenomena like suppletion (e.g. the past tense of *go* in English being the morphophonologically unrelated form *went*). These stems and forms get special treatment when compiled into the FST, as they are completely hard-coded into the model. While there is limited irregularity in Ojibwe, so we are not currently using this functionality in the *OjibweMorph* model, this functionality would be crucial for other languages that have a greater number of irregular inflectional forms.

Our system allows for inflection to be organized into spreadsheets however the user sees fit. We primarily organize the Ojibwe paradigms according to part-of-speech like VTA (transitive animate verbs) and NID (inanimate dependent nouns). Since this could result in very large spreadsheets especially for verbs which have extensive morphological paradigms, we subdivide files along other dimensions where appropriate. For example, `OjibweMorph` contains a spreadsheet `VTA_IND.csv` which contains the independent forms of the various inflection classes for Ojibwe transitive animate verbs. Conjunct and imperative VTA forms have their own spreadsheets. This is merely a convention that should be tuned to enhance maintainability. In principle, all forms could be listed in a single spreadsheet if so desired.

5.1.2 Nominal inflection and rules

As reviewed in Section 3.4.1, nouns in Ojibwe take inflection related to a variety of different grammatical functions including number, obviation, the diminutive, the pejorative, and

Class	Stem	PersPoss.	Dim.	Poss	Basic.	Form1Surface	Form1Split
NA_C	zhiishiib	-	-	-	ProxPl	zhiishiibag	≪zhiishiib≫ag
NA_C	zhiishiib	1Sg	-	Poss	ProxSg	ninzhiishiibim	ni≪zhiishiib≫im
NA_Cw	mitigw2	-	-	-	Loc	mitigong	≪mitigw2≫ong
NA_Cw	mitigw2	-	Dim	-	ObvSg	mitigoonsan	≪mitigw2≫oonsan

Fig. 2 Sample of a nominal spreadsheet with only an immediately relevant subset of columns for this toy example. The actual spreadsheets also include columns for Paradigm, Lemma, the pejorative (Pej), preterit (Pret), Form1Source, and additional trios of Form columns.

possession. Here, we detail our approach to encoding this morphology into the FST via morphological spreadsheets, and we will refer to the example in Figure 2 throughout. Beyond the core columns reviewed in 5.1.1, the noun spreadsheets use the following columns (with naming conventions primarily supplied by those established within the OPD):

- **PersPoss**: The person, number, and obviation of the possessor (if there is one).
- **Dim**: Whether or not the form is marked for the diminutive.
- **Poss**: Whether or not the form is marked for the special possessive suffix (optionally occurs if there is a value for PersPoss).
- **Pej**: Whether or not the noun is marked for the pejorative.
- **Pret**: Whether or not the noun is marked for the preterit.
- **Basic**: Whether the noun is marked for animacy/obviation/number, locative marking, or vocative marking.

For purely organizational purposes, we have separated spreadsheets by phonological stem class. For example, there are separate spreadsheets for NA_C (animate noun stems that end in a consonant) versus NA_Cw (animate noun stems that end in a consonant followed by a "w"). This keeps the number of forms within a given spreadsheet at a manageable amount, and is a logical split for those familiar with the language. All spreadsheets within a given paradigm have the same number of rows, representing all of the possible combinations of nominal inflection including stacked morphemes. For example, animate independent nouns (na) have 440 possible forms, while animate dependent nouns (nad) have 408. This difference is due to the fact that nads cannot occur without a possessor, so can only be marked for a subset of the possible forms that nas can. Similarly, inanimate independent nouns (ni) have 306 possible forms, and inanimate dependent nouns (nid) 286. There are fewer inanimate forms compared to animate because there is no contrast in obviation on inanimate nouns.

The tagset for all nouns has the same basic maximal format, given in (17), along with a handful of example `form:tag` pairs.Again, we have chosen tags based on those used in the OPD, which are in line with those generally used by scholars and students of Ojibwe, to maximize usability for our primary user base.

(17) `Lemma+Paradigm+Dim+Poss+Pej+Pret+Basic+PersPoss`

a. `ninzhiishiibensimishag:zhiishiib+NA+Dim+Poss+Pej+ProxPl+1SgPoss`
b. `nimaamaayinaaban:maamaa+NAD+Pret+ProxSg+ExclPoss`
c. `ishkodeng:ishkode+NI+Loc`

As can be seen in the examples, not all of the tag slots are always filled. For example, if a noun is not possessed, as in (17-c), there will be no value for PersPoss. The tag values are taken directly from the spreadsheet, so can be modified there as needed.

Let us turn now to some of the specifics of how the morphophonology is encoded in the spreadsheets and *xfst* rules with our examples in Figure 2. The discussion is not meant to be exhaustive, but a general overview of the approach. Full details can be found in the documentation within the `OjibweMorph` GitHub repository.

To get the discussion off the ground, first, notice the presence of an underlying symbol "w2" in Figure 2, which appears at the end of the NA_Cw stem *mitigw*. All stems of this class end in this special "w". This is necessary because not all "w"s behave the same way with respect to the phonological rules. We made use of a number of these underlying symbols in order to ensure the rules are properly targeted and circumscribed, though we used them as judiciously as possible to avoid over-complicating and making the inflection less readable to non-experts.

Class	Stem	Order	Subj	Obj	Mode	Neg	Form1Surface	Form1Split
VTA_n	miin1	Ind	2SgSubj	1SgObj	Prt	Neg	gimiizhisiinaaban	gi≪miin1≫i1siinaaban
VTA_n	miin1	Ind	2SgSubj	3SgObvObj	Prt	Neg	gimiinimaasiibaniin	gi≪miin1≫imaasiibaniin

Fig. 3 Example of a verbal spreadsheet with only the immediately relevant subset of columns. The actual spreadsheets additionally include columns for Paradigm, Lemma, Head, Form1Source, and additional trios of Form columns.

Second, notice that there are a number of differences between the inflection in the split column and how it actually appears in the surface form. They are as follows:

- The underlying symbol "w2" does not appear in the surface form (in the sample shown here)
- The first person prefix with the underlying form *ni* surfaces as *nin* when attaching to *zhiishiib*

Both of these divergences from the underlying form in the surface form can be captured by general rules that apply across the entire language. We state these rules in non-formal terms, but in the model they are formalized through *lexc* re-write rules:

- w2Deletion (simplified): w2 deletes word finally, before consonants and vowels "o" and "oo" (so remains before vowels "a", "aa", "i", "ii", and "e").
- PrefixNInsertion: insert "n" after the person prefix "ni" when the sound to the right is "d", "j", "z", "zh", or "g".

The w2Deletion rule represents a typical case for our rules: a rule that targets a change in the stem.[4] The PrefixNInsertion rule has a slightly different motivation, but does represent what we think is the simplest possible solution to the observed allomorphy, where their form is dependent on the phonological properties of whatever is to the right. This is often the stem, but can also be prenouns (and, for verbs, preverbs). We discuss in detail how prenouns and preverbs are implemented in the FST below, but for now consider the example in (18), where the noun *mashkiki* "medicine" is modified by the lexical prenoun *maji-* "bad" and possessed by a third person singular. We can see the prenoun comes between the stem and the person prefix.

(18) *omaji-mashkiki*
o-maji-mashkiki
3-bad-medicine
"his/her bad medicine"

The upshot is that the allomorphy of the person prefixes cannot be reduced to the phonological properties of the left edge of stems, since other elements (i.e. preverbs or prenouns) can always intervene, creating a different phonological context. As such, they are not naturally accounted for within the spreadsheets (which can only capture stem-conditioned allomorphy on the inflection). Instead, re-write rules are a simpler and more flexible means to capture this allomorphy, since they can apply regardless of whether the element to the right of the person prefix is a stem or a preverb/prenoun.

5.1.3 Verbal inflection and rules

The principles and basic formatting of the verbal inflectional spreadsheets and rules are identical to that described in the previous section for nouns. A toy example is given in Figure 3. The main difference between the two is in the specific columns we made use of to indicate the grammatical function of the different inflectional patterns. Besides the core columns described in Section 5.1.1, the verb spreadsheets make use of the following:

- **Head**: The person, number, animacy, and obviation features of the relative clause head (participle order only).
- **Subject**: The person, number, animacy, and obviation features of subject.

[4] In general, it is possible to have an FST that avoids even these sorts of rules. In these cases, one must truncate the stem so that it only includes characters that do not alternate, and expand the "inflection" to include bits of the stem that undergo alternations. An example of this is the FST produced for East Cree by Arppe, Junker, and Torkornoo (2017). While our approach requires rules, it has the advantage of using stems as listed in the OPD, which is the standard in the community, and we therefore avoid creating bespoke stem forms specific to the FST.

- **Object**: The person, number, animacy, and obviation features of object (transitive paradigms only).
- **Mode**: Algonquianist categories related to the epistemic/modal/aspectual system.
- **Negation**: Whether the verb is positive or negative.

As with the nouns, there is a set order for the tagset on verbs, with conventions being based in the Algonquianist tradition of grammatical description. We give this in (19) with a handful of examples.

(19) `Lemma+Paradigm+Order+Negation+Mode+Subject+Object+Head`

a. `zanagad:zanagad+VII+Ind+Pos+Neu+0SgSubj`
b. `waabamaasiwagwaa:waabam+VTA+Cnj+Neg+Neu+1SgSubj+3PlProxObj`
c. `nebaajig:nibaa+VAI+Pcp+Pos+Neu+3PlProxSubj+3PlProxHead`
d. `miijidaa:miijin+VTI+Imp+Sim+InclSubj+0SgObj`

To keep the total number of spreadsheets manageable, we have organized them into separate files based on paradigm and order, with each one containing the possible different stem classes for the relevant combination. As with the nouns, these stem classes ensure that proper allomorphs of the inflection are specified for a given verb, since we aim to minimally rely on the use of phonological rules to derive stem-conditioned allomorphy of the inflection.

Let us go through the examples in Figure 3 to reinforce the approach. First, notice the stem *miin1* "give it to him/her" ends with the underlying symbol "n1", which was referenced in Section 5.1.1. This is the "changeable n" that palatalizes in certain conditions. Specifically, we see this palatalization when the suffixal inflection begins with the special theme sign "i1", which indicates the object is a first person, but not with a regular "i". This underlying difference between the special "i1" and "i" is seen, respectively, in the pair of surface forms *gimiizhisiinaaban* versus *gimiinimaasiibaniin*. We can capture this with the following rules:

- N1RULE: Stems ending in "n1" palatalize to "zh" when the suffix complex starts with the first person theme sign "i1".
- DEFAULTRULE: After all other rules have applied, turn "i1" to "i" and "n1" to "n".

The rules will give us the correct surface form of *gimiizhisiinaaban*, where first "i1" triggers "n1" to palatalize to "zh" (N1RULE), then "i1" is turned to "i" (DEFAULTRULE). For *gimiinimaasiibaniin*, the N1RULE does not apply since there is no special "i1" to trigger it, so the DEFAULTRULE turns "n1" to "n" to get the correct surface form.

5.2 Prenouns and preverbs

As previewed briefly in Section 5.1.2, prenouns and preverbs are semi-phonologically dependent elements that are added before a verb stem, but after the prefixal inflection (if there is any). Together with the stem and the inflection, preverbs form a single prosodic and orthographic word often referred to as the *verbal* or *nominal complex*. There is overlap between the preverbs and prenouns, but preverbs are more common and have more varied categories, so we will focus the discussion there, pointing out where generalizations are extended to the prenouns.[5]

There are five basic types of preverbs and a template that dictates the order in which they combine. SUBORDINATING preverbs occur first and form structures such as relative clauses. TENSE preverbs are second, and can stack together such that there is more than one tense preverb within a clause. DIRECTIONAL preverbs, which indicate the motion of travel through time or space are third. RELATIVE preverbs are fourth and serve to relate the verb to the time, place, or reason for the event predicated by the verb. Last are LEXICAL preverbs, which are often further broken down into sub-categories like "manner", "quality", and "quantificational" preverbs. It is mainly lexical preverbs that also have a double life as prenouns. Multiple lexical preverbs can be stacked, and there are weak ordering restrictions similar to the intuition in English that "three big red cars" is better than "three red big cars".

[5]There are also cases where preverbs/prenouns can be attached to adverbs. For example, *mewinzha* "long ago" can be modified with the intensifying preverb *gichi-* resulting in *gichi-mewinzha* "very long ago". This too is handled in our FST by allowing lexical preverbs to attach to adverbs.

Form	Tag	InputParadigm	InputClass	OutputParadigm	OutputClass
dizo	Reflex/dizo	VTA	VTA_C	VAI	VAI_rfx
idizo	Reflex/dizo	VTA	VTA_n	VAI	VAI_rfx

Fig. 4 Example of derivational morphology spreadsheets.

A nearly maximal example that has every type of preverb from the template except a lexical preverb can be found in (20) (example taken from the OPD; translation original to the present paper). From left to right, first we have the subordinating preverb *gaa-*, which forms a relative clause. Second, the tense preverb *gii-*, which marks the past. Third, the directional preverb *pi-*, which in this case indicates movement towards the speaker (so the movement indicated is "coming" versus "going"). Fourth, there is the relative preverb *onji-*, which indicates the action of the verb is located in space (as opposed to, for example, time). Finally, there is the verb itself, *ayaa* "be", which is inflected with second singular subject conjunct morphology *-yan* (with neutral mode and positive polarity).

(20) *gaa-gii-pi-onji-ayaayan*
gaa-gii-pi-onji-ayaa-yan
REL-PAST-HITHER-SPACE-be-2Sg
"(The place) where you came from"

While examples like this, where four preverbs are stacked, are not the most common occurrence, having one or two preverbs is fairly typical. Tense preverbs are especially common. Preverbs (and prenouns) are always spelled with a hyphen separating them from the main stem.

If the inflection includes a person prefix (see again the example in (18) with the third person prefix *o-*), as is the case with possessed nouns and many independent order verb forms, the inflectional prefix appears as far to the left as possible. That is, the person prefix (if present) is always the initial element and to the left of all of the preverbs or prenouns. The person prefix is also always spelled without a hyphen between it and whatever is to the right of it, be it the stem or a prenoun/preverb.

In `OjibweMorph`, we have a number of template spreadsheets that represent the different types of preverbs described above. The technical challenge that preverbs present is twofold: (i) we aim to capture the major ordering restrictions, so cannot just take a "bag-of-preverbs" approach where they freely combine, and (ii) preverbs must be *inserted* between the prefixal inflection as specified in the spreadsheets and the verb stem. While ordering restrictions can easily be handled using *lexc* continuation lexica, insertion between the person prefix and verb stem requires the use of flag diacritics to ensure that person prefixes are combined correctly with verb forms. Both of these are handled automatically during compilation. Sometimes additional flag diacritics are needed in the preverb template code to model finer distinctions. For example, subordinating preverbs can only attach to conjunct forms. The user can easily model such restrictions by manually adding flag diacritics into the template code.

5.3 Derivational morphology

The next morphological component is a spreadsheet specifying derivational morphemes. Derivational morphemes are affixes (in Ojibwe, always suffixes) that can create new stems with some degree of productivity. So far, we have only modeled a small handful of the most productive derivational processes, but have created a framework that can eventually be extended to capture the full known range.

The example spreadsheet in Figure 4 shows an illustrative sample of our derivational spreadsheet. The following columns are used:

- Form: Underlying orthographic form of the morpheme.
- Tag: Tag associated with the morpheme on the "upper" side of the FST.
- InputParadigm: Paradigm tag for the type of stem the morpheme can attach to.
- InputClass: Stem class tag for the type of stem the morpheme can attach to.
- OutputParadigm: Paradigm tag for the type of stem the morpheme should produce.
- OutputClass: Stem class tag for the type of stem the morpheme should produce.

Full_Form	Clitic_Form	POS
dash	sh	AdvConj

Fig. 5 Example of enclitic spreadsheets.

In the example in Figure 4, we see there are two different forms for the REFLEXIVE morpheme, which can take a VTA_C stem (the input class) such as *dazhim* "talk about him/her" and make it reflexive *dazhindizo* "talk about himself/herself". Setting aside the form for the moment, the resulting stem from this process is a special class of VAI labeled VAI_rfx. Specifying this output class allows these stems to be inflected in the proper manner. Specifically, reflexive verbs in Ojibwe inflect with the same inventory of morphology as a VAI, but differ in that they have an (implicit) object that always matches the person, number, and obviation of the subject. They are therefore handled as a stem class unto themselves with the inflection separated into a unique spreadsheet.

Turning to the allomorphy and rules, in the case of *dazhim*, the form of the verb becomes *dazhindizo* by suffixing the form *-dizo* to the end of the input stem. Note the change of the stem-final "m" to "n". This is the result of a general phonological rule called NASALASSIMILATION, which changes the nasal consonant "m" to "n" when the sound to the right is "z", "g", or "d". Since the reflexive morpheme starts with "d", this rule is triggered to produce the correct surface form. Compare this to *-idizo*, the form for the VTA_n stems. Recall these stems end with the underlying and abstract symbol "n1", as in *jiichiigibin1* "scratch him/her". The derived reflexive takes the form *jiichiigibinidizo* "scratch himself/herself", where the only change from the underlying to surface form is the application of the DEFAULTRULE, which converts the underlying symbol "n1" to "n".

5.4 Clitics

The final morphological component handles *clitics*, morphemes that are syntactically independent, but phonologically dependent on another word. Clitics are generally not selective about the part-of-speech of their "host"—that is, they can attach to a wide variety of different types of words (e.g. Zwicky & Pullum, 1983). This lack of pickiness in selecting a host is the fundamental motivation for creating a separate module. For example, derivational morphemes apply to a specific paradigm and/or inflectional class—clitics do not show such fine-grained distinctions, and can in principle attach to any type of word.

Clitics can be further categorized by whether they attach to the front of their host (proclitics) or the end of their host (enclitics). In our approach, separate spreadsheets are used for these two types of clitics. In Ojibwe, there are only enclitics, so we describe the spreadsheets relative to that sub-type. Everything we say applies to the proclitics as well, except for the fact that the end result would attach the clitic to the left edge of a word rather than the right.

The format of our clitic spreadsheets is exemplified in Figure 5. There are just three columns:

- Full_Form: The full form of the element (if relevant), or dictionary citation form.
- Clitic_Form: The reduced (clitic) form of the element.
- POS: The part-of-speech of the clitic.

In the example spreadsheet here we have just a single clitic with the full form *dash*, spelled as an independent word, and the reduced form *sh*. This element is classified as a conjunctive adverb (POS taken from the OPD). It roughly translates to "but" in English, but its actual meaning and distribution is not precisely equivalent. More technically, it is a "second position" discourse marker (Fairbanks, 2016). That is, it always appears immediately following the first word of the sentence, regardless of the category of that first word. In the written language, clitics are distinguished from affixes by being separated from their host with a hyphen, as shown in the sentence in (21), taken from the OPD.

(21) *Gaawiin-sh niminjimendanziin*
gaawiin=sh ni-minjimendan-ziin
NEG-but 1-remember-NEG
"But I don't remember it"

Lemma	Stem	Paradigm	Class	Translation	Source
miizh	miin1	VTA	VTA_n	give (it) to h/	https://ojibwe.lib.umn.edu/main-entry/miizh-vta
nibaa	nibaa	VAI	VAI_VV	s/he sleeps, is asleep	https://ojibwe.lib.umn.edu/main-entry/nibaa-vai
inini	ininiw2	NA	NA_Vw	a man	https://ojibwe.lib.umn.edu/main-entry/inini-na
jiimaan	jiimaan	NI	NI_C	a boat; a canoe	https://ojibwe.lib.umn.edu/main-entry/jiimaan-ni

Fig. 6 Example of lexical spreadsheets.

The end result is that our FST will recognize any word that has *-sh* attached to the right edge — this includes all nouns and verbs after any derivational and inflectional morphology has been applied, as well as all other parts-of-speech. The resulting analysis adds the tag `+CL/ADVConj/dash` at the right edge of the tagset. For example, the word *gaawiin-sh* in (21) receives the analysis `gaawiin+ADVNeg+CL/ADVConj/dash`.

6 OjibweLexicon: Lexical Module

`OjibweLexicon` is a repository that houses a lexical database of Ojibwe words that can be integrated into the *OjibweMorph* FST. This repository represents the culmination of both decades of lexicography on the Ojibwe language, as well as the end-state of the work our team did to process lexical data from a number of different sources. As described in Section 4.2, the primary source of lexical data is the Ojibwe People's Dictionary (OPD), but we have also started the process of integrating data from other sources such as book glossaries and the fieldwork notes of the team. In this section, we describe the process we undertook to structure and integrate lexical data.

6.1 Data structure

As with the morphological module, our lexical database is structured using *CSV* files to allow for both human and machine readability. A sample spreadsheet is shown in Figure 6. There are six columns, where each row of the spreadsheet is a unique lexical item:

- Lemma: Dictionary citation form.
- Stem: Underlying phonological form, often including abstract underlying symbols.
- Paradigm: High-level part-of-speech tag.
- Class: Phonological stem class.
- Translation: Meaning of the lemma in English.
- Source: URL or other code to indicate where item comes from.

We have divided the spreadsheets by high-level part-of-speech (based on the ones typical within the Algonquianist tradition), so there are separate files for verbs, nouns, adverbs, pronouns, numerals, and so on. This division by part-of-speech is essential for the compilation process into *lexc* code because different parts-of-speech result in very different lexical structures. We additionally use finer divisions for large classes. For example, verbs are divided into *CSV* files along the dimensions of transitivity, animacy, and paradigm (independent/conjunct/imperative). However, this finer division is not essential for compilation. It simply improves readability and organization.

One important remark is the necessity of correctly specifying the phonological stem class of nouns and verbs, and in the general case any elements that inflect within our system. We have designed our spreadsheets such that, in the eventual compilation of the FST, there are distinct continuation lexica according to the class of the stem. Returning to our perennial example, if the stem *mitig* "a tree" was not correctly specified (to focus in just on the animate version of the stem) as NA_Cw, we might incorrectly affix the plural marker *-ag* (characteristic of NA_C stems) rather than the correct allomorph *-oog*. We discuss our approach to automating the classification of nouns and verbs in Section 6.2.

Having consistent and correct underlying forms of the stem is equally important. Consider the form of the lemma *miizh*, which has the underlying form *miin1* with the underlying symbol "n1" (introduced in Section 5.1.3). Even if we correctly specified the class of this verb as VTA_n, we still need to ensure that these stems end with "n1" to ensure our phonological rules are triggered in the proper places.

6.2 Integrating data from the OPD

As detailed in Section 4.2, our main source of lexical data is the Ojibwe People's Dictionary (Nichols, 2012). Our starting point was a raw, but well-structured, database of the dictionary data. We focus here on the aspects of the OPD database relevant to the `OjibweLexicon`, but we will have more to say about other aspects of the database content in our discussion of the model performance evaluation (Section 8).

Like most dictionaries, the OPD is structured as a series of lexical entries comprised of key information about its meaning and use. Each entry has its own unique and standardized URL. Entries are headed by a lemma (the dictionary citation form) which represents the most minimally inflected standalone form that a given item can take. For example, with nouns, this is generally the singular form. Within an entry, the key pieces of information are as follows:

- LEMMA: Dictionary citation form. Unchanged when integrated into `OjibweLexicon`.
- STEM: Underlying phonological form, sometimes including underlying symbols. Adapted to create the STEM for `OjibweLexicon`.
- POS: Part-of-speech tag. Adapted to create PARADIGM tag for `OjibweLexicon`.[6]
- DEFINITION: Redacted by request of OPD Editor.
- FORMS: Key inflectional forms.

For example:

(22) Key data from OPD for lexical item *adik*.
LEMMA: adik
STEM: /adikw-/
POS: NA
DEFINITION: a caribou, a reindeer (redacted in actual dataset)
FORMS: **adik** sg; **adikwag** pl; **adikoons** dim; **adikosh** pej; **adikong** loc

There are two key challenges that we faced when integrating the OPD data into the model that we will consider in turn: (i) the lack of stem class, and (ii) the need to modify and add underlying symbols to stems.

First, the OPD does not currently specify the stem class of a given noun or verb, but, as mentioned above, this is necessary to affix the correct allomorphic forms of the suffixal inflection within our model. It was therefore necessary for us to classify every noun and verb stem from the dictionary. Fortunately, the dictionary is designed such that the stem class of a given lexical item is implicit within the entry. For verbs, it is deducible from the POS, the form of the lemma, and the form of the stem. For nouns, we additionally needed the form of the plural and/or locative inflection, which as shown in (22) is always specified within the FORMS field of the entry. We automated this process by creating a mapping file and script that runs through each lexical entry and determines its class by matching based on its POS, the right edge of the lemma and/or stem, and, if necessary, the patterns of inflection of the key forms in the entry.

Turning now to the second challenge, while there are "special characters" within the OPD that serve a similar function to our underlying symbols, they do not occur in all the same places that we require them, and they do not follow the same conventions as we aimed to follow. For example, the OPD encodes the special "changeable n" that palatalizes to "zh" under certain conditions with a capital "N". Rather than relying on case, we chose to transliterate this character into "n1". These changes and additions were automated using a python script that can map and add these underlying symbols as required.

A final note: For just the data from the Ojibwe People's Dictionary, we have redacted the translation from the public-facing lexical spreadsheets. This was requested by the editors of the dictionary to prevent easy duping of the dictionary. As the translations are not used in our model, this has no impact on our ability to build a robust FST.

[6] A full list and description of these tags can be found here: https://ojibwe.lib.umn.edu/help/ojibwe-parts-of-speech.

7 FSTmorph: Compilation Module

Spreadsheets and lexical data which are housed in the `OjibweMorph` and `OjibweLexicon` repositories are combined and compiled into a finite-state lexicon using program code in the `FSTmorph` repository. We have also released this code as an easily installable Python package (https://pypi.org/project/FSTmorph/). This code is designed to be language-neutral, such that alternative morphological and lexical databases can be swapped in to create an FST for a new language. We reserve the description of the technical details of this compilation to the documentation within the repository to focus on the guiding principles and high-level structure here. Abstractly, compilation encompasses two stages:

1. First, the morphological paradigms stored in `OjibweMorph` are compiled into skeleton *lexc* files, where each inflection class like `VTA_aw` (VTA stems ending in "aw") and `VTA_n` (VTA stems ending in an "n1") contain a single example lexeme with its prefixes, stem and suffixes.
2. Then, inflection classes are populated with lexemes from the `OjibweLexicon` repository (altogether tens of thousands of lexemes may be added in this phase).

Automatic compilation handles many of the trickier aspects of construction of an FST description: The code systematically splits prefixes, stems and suffixes into *lexc* sublexica. It also handles long-range dependencies between inflectional prefixes and suffixes by inserting flag diacritics where needed. Furthermore, it allows for addition of prenouns and preverbs between the person prefix and stem, all the while ensuring that the correct prefixes, stems and suffixes are combined.

By design, compilation is very flexible and maximally independent from the specifics of the target language and its morphological description. A configuration file housed in `OjibweMorph` is used to specify which of the morphological features (like `Mode` and `Order`) provided in the morphological spreadsheets are meant to be included in word analyses (as opposed to simply being present for organization purposes). The configuration file is also used to to distinguish between regular morphological paradigms which apply to multiple lexemes and irregular paradigms for individual lexemes having idiosyncratic inflection patterns. To close our presentation, we illustrate an example of the end result of the compilation pipeline for the word *niwaabamaabaniig* roughly translating to "I was seeing them". This form corresponds to the analysis `waabam+VTA+Ind+Pos+Prt+1SgSubj+3PlProxObj`. In what follows, we walk through a relevant snippet of the *lexc* code generated by `FSTmorph` with our inflectional spreadsheets and lexical sources. While for most users, they will never have to engage directly with the *lexc* code, it is useful to show how our spreadsheets are transformed into code that can be compiled into an FST. In what follows, we assume some knowledge of the *lexc* formalism. For the uninitiated, we recommend first reading Mans Hulden's tutorial on morphological analysis with FSTs (https://fomafst.github.io/morphtut.html). Our example is given from the perspective of generating a form from an analysis.

As with all FSTs, generation starts in the `Root` lexicon, located in the `root.lexc` file within our generated code. This then leads to a variety of word-class specific continuation lexica. Here, we show just the `VerbRoot` lexicon, which leads to paradigm-specific prefix lexica—in this case `VTA_Prefix`. We use the flag diacritic `P.Paradigm.VTA` to mark this form as a VTA. This flag is necessary, because the preverb lexicon (not shown in this example) is shared between multiple different verb paradigms. This flag ensures that we re-enter the correct stem lexicon down the line. All verb-specific lexica that follow are located in the generated file `ojibwe_verbs.lexc`.

```
LEXICON Root
VerbRoot ;

LEXICON VerbRoot
P.Paradigm.VTA VTA_Prefix ;
```

The `VTA_Prefix` lexicon contains the possible prefixes for VTAs—here we just show the piece corresponding to *ni-* in our target form, but all prefixes specified within the `OjibweMorph`

spreadsheets are present in the full model. We also use the flag diacritic `P.Prefix.NI` to eventually connect the prefix with its corresponding set of suffixes later on. We then continue to the `VTA_PrefixBoundary` lexicon, where we mark the boundary between the prefix and the stem with "<<" and continue to the `VerbStems` lexicon. Note that, in our full FST, this is the point where it is possible to branch into the set of preverb lexica—we set that aside for the purposes of this illustration.

```
LEXICON VTA_Prefix
P.Prefix.NI:P.Prefix.NIni VTA_PrefixBoundary ;

LEXICON VTA_PrefixBoundary
0:%<%< VerbStems ;
```

The next three lexica do the following: `VerbStems` inspects the flag diacritics related to Paradigm and ensures that the correct stem lexicon is entered—in this case the `VTA_Stems` lexicon. In the full model, this houses thousands of `lemma:stem` pairs pulled from `OjibweLexicon` along with the corresponding inflectional class. For any given pair, we then continue to a set of lexica specific to the inflectional class of that stem. In this case, we generate the form *waabam*, a VTA_C stem. This leads us through `VTA_Class=VTA_C_Boundary`, where we add the boundary ">>" to mark the right edge of the stem, and head to `VTA_C_Flags`.

```
LEXICON VerbStems
R.Paradigm.VTA VTA_Stems ;

LEXICON VTA_Stems
waabam:waabam VTA_Class=VTA_C_Boundary ;

LEXICON VTA_Class=VTA_C_Boundary
0:%>%> VTA_Class=VTA_C_Flags ;
```

`VTA_C_Flags` brings us into the suffixal inflection. Starting with `VTA_Class=VTA_C_Flags`, we check the prefix flag to ensure we enter the correct lexicon relative to the what person prefix was specified. In the full model, there is the possibility of heading down different paths depending on the the setting of this flag—here we simplify and show the continuation just to `VTA_Class=VTA_C_Prefix=NI_Order=Ind_Endings` corresponding to having specified the *ni-* person prefix. This is our terminal lexicon within our example (in the full model, there is the possibility of heading to the `enclitic.lexc` continuation lexica). This contains the pair that adds the suffixal form *-aabaniig* to the end of the stem, corresponding to the analysis `+VTA+Ind+Pos+Prt+1SgSubj+3PlProxObj`. In the full model there are thousands of such pairs, all pulled from the `OjibweMorph` spreadsheets. The symbol # then marks the end state.

```
LEXICON VTA_Class=VTA_C_Flags
R.Prefix.NI VTA_Class=VTA_C_Prefix=NI_Order=Ind_Ending ;

LEXICON VTA_Class=VTA_C_Prefix=NI_Order=Ind_Endings
+VTA+Ind+Pos+Prt+1SgSubj+3PlProxObj:aabaniig # ;
```

The end result of the above code is the generation of the form *ni<<waabam>>aabaniig*. In the full FST, our phonological rules then apply to delete the stem boundary markers, resulting in the proper surface form *niwaabamaabaniig* for the analysis `waabam+VTA+Ind+Pos+Prt+1SgSubj+3PlProxObj`.

8 Evaluation

In this section, we present a set of tests that evaluated the performance and accuracy of the current version of the *OjibweMorph* FST.

Before proceeding, we would like to emphasize that these results are simply a snapshot of the model's performance at a single point in development. While the model is mature, it is not static, and unless otherwise noted we have not tuned the model to address the gaps that these tests reveal. The tests therefore give something akin to "zero-shot" performance. We will continue to add forms and lexemes as needed to improve coverage and performance based on the results of these tests. As such, users will find that the model is likely performing better than is reported here. The latest values for these tests are reported in the readme documents in the `OjibweMorph` repository, and are dynamically updated along with the model.

8.1 Spreadsheet-based tests

All of our inflectional spreadsheets include both the expected surface form with an example lemma and underlying "split" form. As summarized in Section 5.1.1, the split form is used to generate sets of continuation lexica that form the basis of the inflection within the FST. The surface form, however, is not used in the FST itself, but is instead used to create a set of tests that ensure the correct forms and analyses are being generated by the FST. In short, we create a set of files in the *YAML* format, with each example target form being paired with an analysis. Each file represents a particular paradigm and inflection class. This provides a set of tests targeting whether our FST is being correctly compiled, and that all of our phonological rules lead to the correct changes to transform the underlying form into the proper surface form (and to analyze a given form with the correct set of tags) across all different noun and verb inflection classes. These tests have been implemented using tools from the GiellaLT infrastructure (Wiechetek et al., 2022), and can be automatically run alongside compilation.

At present, our model provides the correct set of analysis tags and generates the correct forms given a tag set for all examples in our spreadsheets. At various points in development of the model, these tests were useful in ensuring that all phonological rules were interacting as expected. We split the tests by paradigm and inflection class, making it easy to see if an issue is relative to a particular morphological class of nouns or verbs. These can also be used as regression tests: if changes to the model are made, we should continue to pass all of these tests. Any backsliding would indicate a bug or other issue with the model. However, these tests only target a small set of example lemmas—at least one in each inflection class. Therefore it provides a broad test that we are correctly generating all of the inflection specified within the spreadsheets, but does not show whether we are correctly generalizing across a wide range of lemmas. Our next set of tests addresses this gap.

8.2 OPD inflectional tests

As previously mentioned in Section 6.2, most entries for nouns and verbs in the OPD contain a set of key inflectional forms along with a tag. There are a total of 75,366 inflectional forms, with 8,565 related to nouns and 66,801 related to verbs. We reformatted these tags to follow the conventions used in our FST, and created a set of tests that allows us to determine whether our FST generates the same analyses as the one provided in the OPD. The inflectional forms listed in each entry were selected by the dictionary editors to (implicitly) show which inflectional class a given stem belongs to. For example, in nouns, we always see the plural form (modulo mass nouns, which cannot be plural), and as necessary the diminutive, pejorative, and locative form. As such, these tests provide a good window into whether we are correctly classifying lemmas into their inflectional classes and paradigms. A misclassification will lead to test failure, thereby flagging any issues in our implementation of this step. For example, if a noun is incorrectly labeled as inanimate, we will produce the wrong plural marker. Many entries also contain "additional forms", which show how a given lemma is inflected for things like tense and possession. These provide additional tests of the inflectional coverage of the model.

The version of the model at the time of writing fails to analyze less than 1% of verb forms (135 total) and less than 1% of noun forms (15 total). In terms of accuracy, verbs have a recall score of 97.01% and a precision score of 77.25%, while nouns have a recall score of 96.92% and a precision score of 83.40%. The high recall score indicates that the FST misses relatively few of the correct analyses (i.e. has a relatively low false negative rate), and the high precision

Speaker	Region	Community	By-Token Failure	By-Type Failure
NJ	Border Lakes	Nigigoonsiminikaaning	5.03% (335/6,651)	7.32% (314/4,285)
GJ	Border Lakes	Lac La Croix	12.32% (9/73)	12.5% (9/72)
ES	Red Lake	Obaashiing	5.65% (539/9,531)	10.51% (518/4,925)
RG	Red Lake	Odaawaa-Zaaga'iganiing	2.54% (56/2,197)	4.44% (55/1,237)
GH	Leech Lake	Jaachaabaaning	2.71% (7/258)	3.39% (7/206)
LW	Leech Lake	Jaachaabaaning	2.63% (5/190)	3.24% (5/154)
LS	Mille Lacs	Aazhomog	8.19% (5/61)	9.61% (5/52)
LSA	Mille Lacs	Lake Lena	3.22% (1/31)	3.44% (1/29)
Unknown	N/A	N/A	0% (0/5)	0% (0/5)
Overall			**5.03% (957/18,997)**	**9.24% (906/9,803)**

Table 5 Failure rates by token and type for each speaker represented in the Ojibwe People's Dictionary (OPD). Region, community, and speaker initials taken from bios on OPD website. These represent the rates at the time of manuscript submission. As we improve the model and the tests, updated metrics will be shared via the GitHub documentation for OjibweMorph.

score indicates that most of the generated analyses are correct (i.e. there is a relatively low rate of false positives). Overall, we can surmise that the model is not severely over- or under-generating—it is usually getting the correct analysis and does not falsely generate incorrect ones.

Our error analysis script further breaks down these values as a function of paradigm and inflectional class, but in the interest of space we present only the overall results. However, we note that these metrics are useful in locating sources of errors. For example, if a particular inflectional class has an especially low score, we can target the parts of the model that encode these forms and debug the issue. We have already made use of this finer-grained error analysis pipeline, going from an initial 15% error rate in the verbs to below 1% in the current version of the model.

We would also like to note that these error rates are inflated compared to the actual accuracy of the model, as many failures arise due to typos in the inflectional forms or tags from the OPD. We are working to fix these typos, but have not yet completed this time-consuming process. Furthermore, there are cases where a form is ambiguous between multiple analyses, but the OPD only indicates one of these possibilities. This impacts the precision scores, since certain valid analyses produced by the FST are not listed within the set of target analyses. These failures are therefore issues with the tests, not with the FST per se. Many errors are also due to known gaps within the model (e.g. inflectional forms that we have not yet added to our spreadsheets), which can be easily remedied as we expand the coverage based on our diagnoses of these errors. For example, we can add an additional form to our spreadsheets, or tweak a phonological rule, and many of the issues will be resolved. We are actively continuing these improvements of the model and the tests as we analyze each error.

8.3 Corpus-based coverage tests

Coverage tests give a sense of how much of a typical text our FST can provide some sort of analysis for. For example, one can calculate the percent of tokens in a corpus that the FST recognizes with some sort of analysis. In this section, we describe and report a set of coverage tests we conducted using a corpus comprised of example sentences published in the OPD.

The OPD corpus contains 7,651 total sentences with 18,997 word tokens and 9,803 unique words, all written in the standard double vowel orthography described in Section 3.3. Sentences are also paired with the original audio, which we made use of here to verify transcriptions. There are sentences from 8 different speakers from communities across the dialect region. This diversity gives us an opportunity to determine the degree to which our model is capturing the variation that occurs within the dialect group.

We first calculated how many words received no analysis by the FST (we refer to these as "failures"). Results are summarized in Table 5. Our overall failure rate was 5.03% by-token and 9.24% by-type. The speaker who contributed the second-most words, NJ, was also the primary consultant for verifying the forms in our inflectional spreadsheets. As expected, these sentences had a relatively low failure rate of 5.03% by-type and 7.32% by token.

In order to understand the types of issues that cause the FST to fail to analyze a given word, we took a randomly generated sample of 100 failures and manually coded them for the

Number of Analyses	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**20**
Count (by-type)	5,469	2,327	524	359	95	57	10	29	0	6	1	11	7	2

Table 6 Distribution of word types from the OPD example sentence corpus that have multiple analyses, given that there was at least one.

source of the failure. Most (61%) were due to a missing lemma. That is, a verb, noun, or other element that is not yet listed within our lexical sources. We also observed that 7% of failures were due to the presence of English (in all cases, English proper names). Both of these failures could be easily fixed by adding these lemmas or English names to our lexical database. In addition, 17% of failures were due to typos in the original sentence from the OPD, for example missing or extra letters. In general, it is straightforward to argue that the model is correct in failing to analyze forms with typos or in languages other than Ojibwe, so 24% of the failures are not *really* failures. Finally, 15% of failures were identified as being due to gaps in our FST other than missing lemmas. For example, not all participle forms have yet been integrated into the FST, and some inflectional variants across the dialect are missing. These failures can largely be resolved by adding additional forms to the inflectional spreadsheets.

If we extrapolate the numbers from our random sample to the full sample, then we estimate a by-type failure rate due to missing inflectional forms of only 2.39%, and a failure rate due to missing lemmas of 5.64%. As we expand our lexicon and refine the inflectional spreadsheets, these errors will be largely eliminated. The flexibility of the approach ensures we are able to easily expand the coverage of the model.

One limitation of the coverage test is we are only measuring whether or not the FST successfully provides at least one analysis—we do not know whether the *correct* analysis is within the set of analyses that are generated, and, if the correct analysis is among the generated analyses, we do not know which one it is. While our inflectional tests presented in the previous section can provide metrics that speak to the model's performance in these areas, we still want to give a general sense of how much ambiguity the FST generates when it does recognize a form. To some degree, this can indicate the precision of the model. We report the distribution of number of analyses for each word type in the corpus in Table 6. The median number of analyses by-type is 1, the modal number of analyses by-type is 1, and the mean by-type is 1.57. This means that most words that get an analysis only receive a single analysis, indicating that the model is not significantly over-generating.

There are a small number of words that have a larger amount of ambiguity. For example, 27 unique words had more than 10 analyses. Taking a look at these forms, all of the ambiguity is expected given a confluence of a small number of factors, which happen to combine to give the appearance of a large degree of ambiguity. For example, *gaa-pimi-ayaamagak* has a twelve-way ambiguity. This is ambiguous first because it could be either a participle or a conjunct order verb. Then, for the ones that are marked as conjunct order, it could either have a singular or plural subject. Finally, there are various ambiguities in the stem/lemma. To take one example, the dictionary lists the lemma *bimi-ayaa* "s/he goes" in its own entry, and so that string is present in the FST as a lemma (generating the analysis `bimi-ayaa+VAI`). However, it can also be formed through an alternative path by combining the directional preverb *bimi-* "along" with the verb *ayaa* "s/he moves" (generating the analysis `PVDir/bimi+ayaa+VAI`).

It is also important to emphasize that just because a word has multiple analyses, that does not mean there is an issue with the FST. Out of context, many words can be interpreted in different ways. That said, most forms are not ambiguous in the context of the sentence they are in. Consider the word *walk* in English. In context, this can unambiguously be a noun (e.g. *I took a walk*) or a verb (e.g. *I love to walk*). However, if we are just analyzing each word token on its own, without attention to context, we should analyze this ambiguously as possibly being a noun or a verb. In other words, we want to generate all possible analyses of a form with the FST. We can then use additional tools such as constraint grammars to determine which analysis within the set is appropriate given the context (e.g. Schmirler, Arppe, Trosterud, & Antonsen, 2018). We are in the process of creating such a program for Ojibwe.

9 Current and future applications

The primary goal of this paper was to describe the development and evaluation of an FST for the Ojibwe language. In this section, we describe the ways in which we are making use of this FST for various applications related to language revitalization and use.

Text analysis

Our first and primary goal in the development of this FST is to support the creation of the first morphologically analyzed text-based corpus for Ojibwe. We are currently in the middle of developing this resource. The initial version will focus primarily on modern texts from the Southwestern dialect group, which will include an estimated 250,000 word tokens in the first release. This will be released on the Korp platform (Borin, Forsberg, & Roxendal, 2012).

We have also made the FST available for use through Dustin Bowers' text analysis tool (https://bowersd.github.io/textAnalysis/). With this tool, a user can input a text and output a table of analyses and links to relevant dictionary entries for each word in the text.

Automatic verb conjugation tool

Our second key application, which also drove a significant amount of the initial development of this model, is an automatic verb conjugation tool for Ojibwe. Users can select a verb, and choose what they want the subject, object, tense, and so on to be, and the application generates the correct form, as well as a table with other related forms. This tool is currently in beta and will soon be more widely released by the *Anishinaabemodaa* language learning platform being developed by the educational technology company CultureFoundry. This conjugation tool will be available to anyone in the world free of charge, but will specifically be used for Ojibwe language instruction at the Kindergarten to Grade 12 levels across many Ontario school boards. Knowing our FST would be used for this tool in particular guided our early choice to create a chunked rather than concatenative model. We wanted to ensure that all forms generated by the model were verified, which is easier to do on our spreadsheet-based approach, where the FST never fills in missing forms in the paradigm with guesses.

Intelligent dictionary search

A third application for which implementation is ongoing is to integrate the FST into the Ojibwe People's Dictionary search bar to create "morphologically intelligent" search for the OPD. This will be similar to what has been created for Plains Cree, available as *itwêwina* (https://itwewina.altlab.app; Arppe et al., 2022). At present, the OPD is not able to handle most inflected word forms, as it only uses a simple string-matching algorithm. As a result, users of the dictionary need to know how to separate inflection from the stem before searching to come up with an accurate result. This poses a barrier for learners and others who may not have that level of meta-linguistic knowledge. With the FST, complex word forms can be analyzed such that any lemmas, preverbs, or other elements that are part of the form can be pulled up in the search, and we can indicate the approximate meaning of the originally searched word by translating the resulting tags into plain English. Using tools within the *GiellaLT* compilation infrastructure (Wiechetek et al., 2022), we also plan to include a set of spelling relaxation rules in this version of the FST, which will allow users to make various errors in their search (e.g. mistake a long vowel for a short vowel), but still come up with a result that most closely matches the input.

Spell-checker

Our fourth application is the creation of a spell-checker for the FST. In general, FSTs can be turned into spell-checkers under the assumption that words with no analysis in the FST are ill-formed in some way and should be corrected. The spell-checker can also provide suggestions for possible forms that may have been intended. These suggestions can be customized and weighted based on common errors, ensuring that the most likely results are prioritized in the set of suggested fixes. We are releasing our spell-checker using the Divvun toolkit (https://divvun.org/) and the *GiellaLT* compilation infrastructure (Wiechetek et al., 2022). Our current model is fully compatible with these toolkits, as it can be compiled with the

Helsinki Finite-State Transducer (*HFST*; Lindén et al., 2011). A beta version of the model is released under the language code `ciw` (https://github.com/giellalt/lang-ciw), and can be installed on Mac and PC using the Divvun installer.

Beyond (Southwestern) Ojibwe

We close this section with a brief discussion of the potential to apply the current toolkit outside of the Southwestern Ojibwe dialect group. We see what we have created as a multipurpose, approachable, and easy-to-maintain set of tools for generating an FST for any morphologically complex language. We are currently working to create a "flattened" version of the Bowers et al. (2017) Odawa parser, where we instead list forms within a set of inflectional spreadsheets, which will increase the maintainability and extensibility of this FST. We are also in discussions with other language communities in the Algonquian family, including various dialects of Cree, on the potential to use our framework to create new FSTs or adapt existing ones to our format. More generally, we are releasing all of our code and data, which we hope can serve as a guide for those who might want to independently create an FST with our toolkit. We aim to provide support to those who are interested in developing these programs for their own language community.

10 Conclusion

This paper detailed the creation, evaluation, and application of *OjibweMorph*, the first morphological parser for the Southwestern Ojibwe dialect group. Our main aim was to detail at a high-level the design choices that went into the creation of this specific FST, as well as the general toolkit we created to aid in the generation of this FST. Our approach consists of three key modules: a set of spreadsheets that define how elements like nouns and verbs can be inflected and how new elements can be derived, a lexical database, and a compilation module. Our goal was to create an easily human-editable format, which can easily be expanded and adapted with minimal technical expertise. We anticipate that both the specific discussion of creating an FST for Ojibwe and the general discussion of our modular approach will be a resource for those who wish to create an FST on a novel language. Finally, the *OjibweMorph* FST will be deployed across a wide range of applications, including academic, educational, and general use tools, showing the broad utility of this type of work for Indigenous and small-footprint languages.

Supplementary information. The code and data related to the work described in this paper can be found in the following three GitHub repositories:

https://github.com/ELF-Lab/FSTmorph
https://github.com/ELF-Lab/OjibweMorph
https://github.com/ELF-Lab/OjibweLexicon

The code and data are publicly viewable and available for use as indicated by the license within each repository.

Acknowledgments. Our most heartfelt thanks and appreciation goes to those speakers of Anishinaabemowin who have shared their language with the Ojibwe People's Dictionary: Eugene Stillday, Gerri Howard, Gordon Jourdain, Leona Wakonabo, Lee Staples, Larry Smallwood, Marlene Stately, Nancy Jones, and Rose Tainter. Nancy Jones deserves a special thanks for aiding in the confirmation of the inflectional forms. We also acknowledge that a citation is not sufficient to recognize the contributions of Dr. John Nichols to this work, who has devoted his life to documenting the Ojibwe language. Thanks to the board of the Midwest Indigenous Immersion Network, and audiences at the University of Alberta and the Algonquian Conference, for feedback. Thanks for Reed Steiner for assisting in the updating of some of the nominal spreadsheets.

Funding. This work was supported by a SSHRC Insight Grant (435-2023-0474) awarded to Hammerly, Arppe, and Silfverberg and a SSHRC Partnership Grant (895-2019-1012) to Arppe, Silfverberg and Hammerly (among others).

Declarations. Hammerly has received research support from CultureFoundry through the Mitacs Accelerate program and an industry research grant at UBC for a project unrelated to the work reported in the present paper. The authors have no other relevant financial or non-financial interests to disclose.

Author Contributions. Conceptualization: Hammerly, Silfverberg; Methodology: Silfverberg, Arppe, Hammerly; Software: Silfverberg, Stacey; Validation: Stacey, Silfverberg, Hammerly; Writing - original draft preparation: Hammerly; Writing - review and editing: Silfverberg, Livesay, Arppe; Funding acquisition: Hammerly, Silfverberg, Arppe; Resources: Livesay, Hammerly.

Ethics. The fieldwork methods described in this paper were approved by the UBC Office of Research Ethics.

References

Arppe, A., Cox, C., Hulden, M., Lachler, J., Moshagen, S.N., Silfverberg, M., Trosterud, T. (2017). Computational modeling of verbs in Dene languages: The case of Tsuut'ina. In *Working papers in Athabaskan (Dene) languages* (pp. 51–69).

Arppe, A., Junker, M.-O., Torkornoo, D. (2017). Converting a comprehensive lexical database into a computational model: The case of East Cree verb inflection. *Proceedings of the 2nd workshop on the use of computational methods in the study of endangered languages* (pp. 52–56).

Arppe, A., Lachler, J., Trosterud, T., Antonsen, L., Moshagen, S.N. (2016). C. Soria, L. Pretorius, T. Declerck, J. Mariani, K. Scannell, & E. Wandl-Vogt (Eds.), *Proceedings of CCURL 2016: Collaboration and Computing for Under-Resourced Languages: Towards an Alliance for Digital Language Diversity* (pp. 1–8). Portorož, Slovenia: European Language Resources Association.

Arppe, A., Poulin, J., Santos, A., Eddie, Neitsch, A., Harrigan, A., Schmirler, K., ... Wolvengrey, A. (2022). Towards a morphologically intelligent and user-friendly on-line dictionary of Plains Cree – next next round. *Presentation at the 54th Algonquian Conference, Boulder, CO.*

Beesley, K.R., & Karttunen, L. (2003). *Finite-state morphology: Xerox tools and techniques.* CSLI, Stanford.

Bloomfield, L. (1957). *Eastern Ojibwa: Grammatical sketch, texts, and word list.* University of Michigan Press.

Borin, L., Forsberg, M., Roxendal, J. (2012). Korp – the corpus infrastructure of Språkbanken. *Proceedings of LREC 2012. Istanbul: ELRA* (p. 474–478).

Bowers, D., Arppe, A., Lachler, J., Moshagen, S., Trosterud, T. (2017). A morphological parser for Odawa. *Proceedings of the 2nd workshop on the use of computational methods in the study of endangered languages* (pp. 1–9).

Chan, V.S.Y., & Hammerly, C. (2025). Evaluating Indigenous language speech synthesis for education: A participatory design workshop on Ojibwe text-to-speech. *Eight workshop on the use of computational methods in the study of endangered languages* (p. 47).

Davis, F., Santos, E.A., Souter, H. (2021). On the computational modelling of Michif verbal morphology. *Proceedings of the 16th conference of the european chapter of the association for computational linguistics: Main volume* (pp. 2631–2636).

Fairbanks, B. (2016). *Ojibwe discourse markers.* University of Nebraska Press.

Forbes, C., Nicolai, G., Silfverberg, M. (2021). An FST morphological analyzer for the Gitksan language. *Proceedings of the 18th sigmorphon workshop on computational research in phonetics, phonology, and morphology* (pp. 188–197).

Fransen, T. (2020). Automatic morphological analysis and interlinking of historical Irish cognate verb forms. In *Morphosyntactic variation in medieval Celtic languages: Corpus-based approaches.* De Gruyter Mouton.

Hammerly, C., Fougère, S., Sierra, G., Parkhill, S., Porteous, H., Quinn, C. (2023). A text-to-speech synthesis system for Border Lakes Ojibwe. *Proceedings of the sixth workshop on the use of computational methods in the study of endangered languages* (pp. 60–65).

Harrigan, A.G., Schmirler, K., Arppe, A., Antonsen, L., Trosterud, T., Wolvengrey, A. (2017). Learning from the computational modelling of Plains Cree verbs. *Morphology*, *27*, 565–598,

Holden, J., Cox, C., Arppe, A. (2022, June). An expanded finite-state transducer for Tsuut'ina verbs. N. Calzolari et al. (Eds.), *Proceedings of the thirteenth language resources and evaluation conference* (pp. 5143–5152). Marseille, France: European Language Resources Association. Retrieved from https://aclanthology.org/2022.lrec-1.551

Hulden, M. (2009). Foma: a finite-state compiler and library. *Proceedings of the demonstrations session at eacl 2009* (pp. 29–32).

Kadlec, D. (2022). *A computational model of Blackfoot noun and verb morphology* (Unpublished master's thesis). University of Lethbridge (Canada).

Kazeminejad, G., Cowell, A., Hulden, M. (2017). Creating lexical resources for polysynthetic languages—the case of Arapaho. *Proceedings of the 2nd workshop on the use of computational methods in the study of endangered languages* (pp. 10–18).

Lachler, J., Antonsen, L., Trosterud, T., Moshagen, S., Arppe, A. (2018). Modeling Northern Haida verb morphology. *Proceedings of the eleventh international conference on language resources and evaluation (lrec 2018).*

Lane, W., & Bird, S. (2019). Towards a robust morphological analyser for Kunwinjku. *17th workshop of the australasian language technology association* (pp. 1–9).

Lindén, K., Axelson, E., Hardwick, S., Pirinen, T.A., Silfverberg, M. (2011). Hfst—framework for compiling and applying morphologies. *Systems and frameworks for computational morphology: Second international workshop, sfcm 2011, zurich, switzerland, august 26, 2011. proceedings 2* (pp. 67–85).

Littell, P., Stewart, D., Davis, F., Pine, A., Kuhn, R. (2024). Gramble: A tabular programming language for collaborative linguistic modeling. *Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation (lrec-coling 2024)* (pp. 7913–7925).

Muradoğlu, S., Evans, N., Suominen, H. (2020). To compress or not to compress? a finite-state approach to Nen verbal morphology. *Proceedings of the 58th annual meeting of the association for computational linguistics: Student research workshop* (pp. 207–213).

Nguyen, M., Hammerly, C., Slifverberg, M. (2025). A hybrid approach to low-resource machine translation for Ojibwe verbs. *Proceedings of the fifth workshop on nlp for indigenous languages of the americas (americasnlp)* (pp. 18–26).

Nichols, J.D. (1980). *Ojibwa morphology* (Unpublished doctoral dissertation). Harvard University.

Nichols, J.D. (2011). *A concise grammar of Minnesota Ojibwe.* (Unpublished Manuscript, University of Minnesota)

Nichols, J.D. (2012). *Ojibwe People's Dictionary.* Retrieved from https://ojibwe.lib.umn.edu

Schmirler, K., Arppe, A., Trosterud, T., Antonsen, L. (2018). Building a constraint grammar parser for plains cree verbs and arguments. *Proceedings of the eleventh international conference on language resources and evaluation (lrec 2018).*

Snoek, C., Thunder, D., Lõo, K., Arppe, A., Lachler, J., Moshagen, S., Trosterud, T. (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* (pp. 34–42).

Steiner, R., & Hammerly, C. (to appear). Refining the phonological analysis of Ojibwe nominal inflection classes. *Proceedings of the 55th algonquian conference.*

Sullivan, M.D. (2016). *Relativization in Ojibwe* (Unpublished doctoral dissertation). University of Minnesota.

Swierzbin, B. (2003). Stress in Border Lakes Ojibwe. H. Wolfart (Ed.), *Papers of the 34th algonquian conference* (pp. 341–370). Winnipeg: University of Manitoba.

Tyers, F., & Castro, S.H. (2023). Towards a finite-state morphological analyser for San Mateo Huave. *Proceedings of the sixth workshop on the use of computational methods in the study of endangered languages* (pp. 30–37).

Valentine, J.R. (1994). *Ojibwe dialect relationships.* The University of Texas at Austin.

Valentine, J.R. (2001). *Nishnaabemwin reference grammar.* University of Toronto Press.

Valentine, J.R. (2024). *Anishinaabemowin.* Retrieved from https://ojibwegrammar.langsci.wisc.edu/

Wang, S., Yang, C., l Parkhill, M., Quinn, C., Hammerly, C., Zhu, J. (2025). Developing multilingual speech synthesis system for Ojibwe, Mi'kmaq, and Maliseet. *Proceedings of the 2025 conference of the nations of the americas chapter of the association for computational linguistics: Human language technologies (volume 2: Short papers)* (pp. 817–826).

Wiechetek, L., Hiovain-Asikainen, K., Mikkelsen, I.L.S., Moshagen, S., Pirinen, F., Trosterud, T., Gaup, B. (2022, June). Unmasking the myth of effortless big data - making an open source multi-lingual infrastructure and building language resources from scratch. *Proceedings of the thirteenth language resources and evaluation conference* (pp. 1167–1177). Marseille, France: European Language Resources Association. Retrieved from https://aclanthology.org/2022.lrec-1.125

Zwicky, A.M., & Pullum, G.K. (1983). Cliticization vs. inflection: English n't. *Language*, 502–513,