A Morphological Parser for Odawa

Dustin Bowers
Antti Arppe
Jordan Lachler
University of Alberta
4-32 Assiniboia Hall, University of Alberta
Edmonton, Alberta, Canada T6G 2E7
dabowers@ualberta.ca
arppe@ualberta.ca
lachler@ualberta.ca

Sjur Moshagen
Trond Trosterud
Department of Language and Culture
The Arctic University of Norway
Box 6050 Langnes
N-9037 Tromsø, Norway
sjur.n.moshagen@uit.no
trond.trosterud@uit.no

Abstract

Language communities and linguists conducting fieldwork often confront a lack of linguistic resources. This dearth can be substantially mitigated with the production of simple technologies. We illustrate the utility and design of a finite state parser, a widespread technology, for the Odawa dialect of Ojibwe (Algonquian, United States and Canada).

1 Credits

We would like to thank Rand Valentine, Mary Ann Corbiere, Alan Corbiere, Sjur Moshagen, Trond Trosterud, Lena Antonsen, Miikka Silfverberg, Ryan Johnson, Katie Schmirler, Sarah Giesbrecht, and Atticus Harrigan for fruitful discussions during the development of this tool. We would also like to thank two anonymous reviewers for their helpful comments. This work was supported by a Partnership Development Grant (890-2013-0047) from the Social Sciences and Humanities Research Council of Canada and a Research Cluster Grant from the Kule Institute for Advanced Study at the University of Alberta.

2 Introduction

Language communities and linguists conducting grammatical or documentary fieldwork often confront a lack of linguistic resources. There may be incomplete prior grammatical descriptions of a language, an absence of written texts, or little infrastructure to help produce them. However, even with very few resources, linguistic technology can be produced to facilitate resource production. Building on prior work (see Trosterud 2005, Snoek et al 2014), the Alberta Language Technology laboratory (ALT-lab[1]) at the University of Alberta has produced a finite state model of the Odawa dialect of Ojibwe (otw, Algonquian, United States and Canada).[2] The production of this tool opens the door to faster editing and grammatical annotation of written texts, increases usability of electronic dictionaries, and provides a simple way to produce and check paradigms. We here summarize key features of the model of Odawa, and highlight some applications in which it is being used.

3 Basics of Finite State Machines

Finite state machines are a popular representation for a simple class of formal languages known as regular languages (Jurafsky and Martin 2000, Beesley and Karttunen 2003). Most importantly for our purposes, the phonology and morphology of natural languages may be productively modeled as regular languages (Johnson 1972, Koskenniemi 1983, Kaplan and Kay 1994). That is, the set of legal words of a natural language can be compactly represented with a finite state machine.

To take a simple example, we illustrate a finite state grammar for parsing the Odawa word for 'sacred stories' *aadsookaanan* into the stem and plural morpheme *aadsookaan-an* (doubled vowels indicate long vowels). Reading from left to right, the first step is to recognize *aadsookaan* 'sacred story' as an existing Odawa noun stem (indeed, a legal word), which in this case is followed by the plural morpheme *-an*, after which the word may not be further inflected and must end. This parse, and any other exhaustive segmentation of the string, is returned. This sequen-

[1] http://altlab.artsrn.ualberta.ca/

[2] A similar model is under development by the Biigtigong Language Project (Kevin Scannell and John Paul Montano), see http://github.com/jpmontano/fst.

tial decision process is straightforwardly mirrored in finite-state machines, where morpheme boundaries in a word correspond to states, and morphemes form the transitions between states. This can be represented as a directed graph, where states are nodes and morphemes are placed on arcs between nodes, with the phonological and syntactic components of the morpheme separated by a colon (e.g. *aadsookaan:N*).[3] The machine starts from a beginning state, labeled '0', and paths corresponding to legal words end in final states, marked with a double boundary. Hence, the finite state machine that parses *aadsookaan-an*, appears in Figure 1.

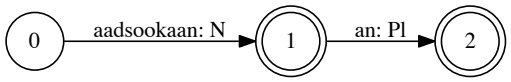

Figure 1: Finite state machine that recognizes the language containing *aadsookaan* and *aadsookaanan*.

4 Odawa Description

Odawa is a dialect of Ojibwe spoken mostly in the Great Lakes region of the United States and Canada. The dialect is endangered, as natural inter-generational transmission of the language ceased during or before the 1960's. Precise numbers of first-language speakers are difficult to come by, though Golla (2007) estimates that there are no more than two thousand speakers. There are efforts to revitalize the language, with immersion and non-immersion programs for children and adults in several communities. The language is predominately written in a romanization called the Fiero double-vowel system. Individual authors vary in whether they explicitly spell phonological processes like word-final devoicing or word-final short vowel lowering.

Ojibwe is a comparatively well-resourced language, as there was a tradition of missionary linguistics dating to the seventeenth century, which culminated in works by Baraga (1878b; 1878a) and Cuoq (1886). Odawa itself has also been the subject of sustained investigation, as seen in the works of Bloomfield (1957), Piggott, Kaye and Tokaichi (1971), Piggott and Kaye (1973), Rhodes (1985a), Valentine (2001), and Corbiere and Valentine (2016).

4.1 Odawa Morphology

Like all Algonquian languages, Odawa has concatenative morphology that is easily represented with finite state machines. Setting aside a full discussion of the inflectional morphology (see Nichols 1980, Valentine 2001), nouns and verbs may be preceded by person prefixes, which may be followed by zero to six adverbial/adjectival prefixes. The full range of available suffixes varies between nouns and sub-classes of verbs, but they all have slots for realizing number, mood/aspect, and a second number slot specifically for arguments indexed by prefixes. Example (1) illustrates these slots for nouns (example from Valentine 2001:207).

(1) g-makko-waa-bn-iin
2-box-2.PL-PRF-PL
'your folks' former boxes'

Nouns have additional slots for diminutive/contemptive suffixes, a pejorative suffix, and an optional possessive suffix. These suffixes are exemplified in the following synthesized example.

(2) g-makk-oons-sh-im-waa-bn-iin
2-box-DIM-PEJ-POS-2.PL-PRF-PL
'your folks' former darn little boxes'

The verbal suffix template includes a position for negation, and in the case of transitive verbs with animate objects, a slot for object agreement (called a 'theme-sign' in the Algonquianist tradition, Brittain 1999, McGinnis 1999). Example (3a) illustrates the major verbal slots with a verb that subcategorizes for an animate object (example synthesized, see Valentine 2001:293), whereas example (3b) shows a verb with two adverbial prefixes ('preverbs' in the Algonquianist tradition, example courtesy of Alan Corbiere).

(3) a. n-waabm-aa-swaa-mnaa-dgen-ag
1-see-3.OBJ-NEG-1.PL-DUB-3.PL
'Perhaps we (excl.) don't see them'

b. n-gii-zhi-bgosenm-aa-naan-ig
1-PST-thus-beg-3.OBJ-1.PL-3.PL
'We (excl.) did thus beg of them'

The examples in (3) illustrate the inflection of verbs in matrix clauses ('independent order'

[3] This makes the machine a transducer rather than a standard automaton.

in Algonquianist terminology). The slot order and phonological spell-out changes significantly in embedded clauses ('conjunct order' in Algonquianist terminology), an example of which is provided in (4). For simplicity, we will henceforth restrict attention to matrix clause forms.

(4) waabm-aa-siiw-angi-dwaa-wenh
see-3.OBJ-NEG-1.PL-3.PL-DUB
'Perhaps we (excl.) don't see them'

4.1.1 Long-Distance Dependencies

As mentioned in the previous section, number suffixes occur in different slots than the affixes for realizing person information. This is most clearly seen in how each person prefix subcategorizes for distinct sets of number suffixes. For instance, the first person prefix *n-* in (3a) is compatible with *-mnaa* '1.PL', but not with the suffix *-waa* '2.PL' seen in (1).

By contrast, the second person prefix *g-* is compatible with both *-mnaa* '1.PL' and *-waa* '2.PL', as seen in (5).[4]

(5) a. g-waabm-aa-mnaa-dgen-ag
2-see-3.OBJ-1.PL-DUB-3.PL
'Perhaps we (incl.) see them'

b. g-waabm-aa-waa-dgen-ag
2-see-3.OBJ-2.PL-DUB-3.PL
'Perhaps you folks see them'

Finally, third person prefixes are not compatible with *-mnaa* '1.PL', but are compatible with *-waa* '2.PL'. This is shown in (6).[5]

(6) w-waabm-aa-waa-dgen-an
3-see-3.OBJ-NEG-2.PL-DUB-3.OBV
'Perhaps they see him'

These interactions between non-adjacent slots are important to track when developing a finite-state model, since finite state machines standardly have no memory, and thus cannot recall what a previous affix was. Without extensions to augment the power of the grammar, discussed in section 5.3.1, these dependencies can result in very large and difficult to maintain machines.

[4] As the examples in this section show, glosses for these number suffixes are somewhat imprecise, as their distribution indicates that their meaning is closer to "plural not including first person" (*-waa*) or "plural including first person" (*-mnaa*).

[5] The attentive reader may notice that the final suffix in (6) is *-an* '3.OBV'. This suffix marks obviation, which is required when two animate third persons are in the same clause. Functionally, obviation serves to allow predications with two third person participants to be expressed despite a general ban on arguments with the same person values.

4.2 Odawa Phonology

Algonquian languages commonly have uncontroversial phonology, with small phoneme inventories, simple syllable structure and few paradigmatic alternations. Much of Odawa phonology is no different, with many processes centering around adjusting consonant clusters or vowel hiatus in typologically unsurprising ways. For instance, the animate plural morpheme *-ag*, as in *daabaan-ag* 'cars', loses its vowel when preceded by a long vowel, as seen in *aamoo-g* 'bees'.

Odawa phonology recently became substantially more intricate. In the early part of the twentieth century, unstressed vowels became radically reduced or deleted (Bloomfield 1957:5). Since stress was assigned by left-to-right iambs, and person prefixes were incorporated into the phonological word, this resulted in paradigmatic alternations like those in (7):

(7)

'shoe'	'my shoe'	
makizin	ni-makizin	UR
(makí)(zín)	(nimá)(kizín)	Stress
(m_kí)(zín)	(n_má)(k_zín)	Deletion
mkizin	nmakzin	SR

The innovation of the unstressed vowel deletion process has triggered an ongoing period of phonological changes, most saliently including the rise of novel prefix allomorphs and a decline in the use of stem-internal alternations (Rhodes 1985b, Bowers 2015:Ch 5). For instance, while *n-makzin* 'my shoe' is still recognized as a legal form of 'shoe', speakers also productively use *ndoo-mkizin* 'my shoe' (Rhodes 1985a). Indeed, person prefixes now appear in a variety of forms that are used interchangeably. In addition to *n-* and *ndoo-* for first person, we also find *nda-* and *ndi-*. Parallel allomorphy is seen for second person and third person prefixes, see Valentine (2001:62) and Bowers (2015:ch 5) for descriptions of how these prefixes arose.

There appear to be other changes as well, since the paradigms elicited from a native speaker of the language by Valentine (2001) often include some forms that cannot be derived by the unstressed vowel deletion process from the hypothesized underlying representations. The examples in (8) illustrate some of the forms that differ from their expected surface values, given Valentine's URs (Valentine 2001:233, 259).

(8) a. ‘you folks arrive’

gi-dagoshin-am	UR
(gidá)(goshí)(nám)	Stress
(g_dá)(g_shí)(nám)	Deletion
gdagshinam	Expected
gdagoshnam	Listed

b. ‘they don’t taste good’

minopogod-sin-oon	UR
(minó)(pogód)(sinóon)	Stress
(m_nó)(p_gód)(s_nóon)	Deletion
mnopgosnoon	ds→s
mnopgosnoon	Expected
mnopgsnoon	Listed

The full scope of these changes is currently being investigated. For the time being, our model implements the prefix changes and paradigm leveling innovations, but still enforces unstressed vowel deletion as if no exceptions had arisen.

5 Odawa Model

5.1 Design considerations

Our finite state model of Odawa has been written in *lexc*, a widely used programming language for writing finite state machines, which is recognized by the *xfst*, *hfst* and *foma* compilers (Beesley and Karttunen 2003, Lindén et al 2011, Huldén 2009). Additional processing of the output of the morphological module is carried out with *xfst* scripts. Our source code may be accessed at `https://victorio.uit.no/langtech/trunk/langs/otw/src/`.

Finite state machines allow morphological structure to be encoded in a variety of ways. As stated in section 3, a natural representation of concatenative morphology maps morpheme slots to states, and morphemes to the labels of arcs between the states. This is not, however, the only possible representation. Developers may choose to treat sequences of affixes, or even whole words, as unanalyzed wholes (often referred to as ‘chunking’). Such an approach may be particularly useful if combinations of morphemes have non-compositional semantics, if segmentation of morphemes is difficult, or if pre-existing resources (like a database) already treat morphology in this way.

Another modeling decision concerns whether one deals with morphophonological alternations at stem-affix junctures by dividing stems into subtypes which are each associated with their own inflectional affix sets that can simply be glued onto the stem, or whether one models such morphophonological alternations using using context-based rewrite rules (roughly of the SPE type), or some combination of these approaches. In our case, we have chosen to model such morphophonological alternations entirely with rewrite rules, thus requiring no stem subtypes but the marking some orthographemes at the stem-suffix juncture with special characters to control the triggering of these rules.

Furthermore, morphosyntactic features need not perfectly mirror the target language. This is especially relevant in languages like Odawa where notionally related features like person and number are realized in possibly multiple disjoint locations, or if the morphological realization of a category like subject or object agreement varies between paradigm types.[6] In such cases, it can be convenient for non-specialist use of the parser, as well as for integration with other software applications, to depart from a close mapping between the target language and the model. See section 5.3.2 for further discussion.

In the case of Odawa, the concatenative, compositional nature of the morphology lends itself to a splitting approach, though a brute-force listing of entire suffix sequences may be attractive to avoid having to deal with the potentially disjoint multiple exponents of e.g. person/number features (cf. Arppe et al., fc). Splitting the morphemes has resulted in a concise and general description, which allows our model to generate inflected forms even if the cell of the paradigm was not enumerated in our source material.

Furthermore, Rand Valentine (p.c.) indicates that Ojibwe dialects often differ not in entire suffix chunks, but in the use of specific suffixes. Avoiding a redundant brute-force listing of suffix sequences thus positions our model to be easily extended to other dialects of Ojibwe, as the individual morpheme changes can be easily identified and carried out.

5.2 Phonology Module

As indicated in section 4.2, the model needs a phonological module to map the representation /gi-makakw-waa-bany-an/ ‘your folks’ former boxes’ (Figure 2) to the actually observed

[6] This is the case in Odawa, where subject and object agreement occur in different morphological slots for verbs in matrix or embedded clauses.

gmakkowaabniin (example 1). We use a cascade of finite state transducers that are formally identical to *SPE*-style phonological rewrite rules (Chomsky and Halle 1968). This phonological module is composed with the morphological model, so that the morphological strings are modified by the phonology until they match surface forms of the language.

5.3 Morphological Module

The morphological module follows the slot structure of the language quite closely. That is, as each morpheme is encountered, a morphological feature is emitted. For instance, the section of the machine that handles example (1) corresponds to Figure (2). Finally, 14,237 lexical entries, drawn from Corbiere and Valentine (2016), make up the lexical content of our model.

5.3.1 Long Distance Modelling

Section 4.1.1 illustrated some of the long-distance dependencies that occur in Odawa. Recall that these relationships can make a standard finite state machine quite large and cumbersome to maintain and develop. The machine can be substantially compressed, at some cost in parsing speed, by introducing limited memory into the model with the flag diacritic extensions in *lexc*. When a morpheme m that interacts non-locally with another morpheme n is encountered, this information is stored in memory. When morpheme n is encountered, the information is retrieved, and if m is compatible with n, the parse continues.

To see this, consider Figure 3, which diagrams the person-number interaction for noun possession. In the figure, stored information is signaled with a flag diacritic of the form *!P.Per.X*, or 'positively set the person feature to X', while accessed information is signaled with *!D.Per.X*, or 'deny if the person feature is X'. Hence, if the first person prefix *ni-* is read, the machine will rule out following it with *-waa* '2.PL', which is incompatible with the first person feature.

5.3.2 Unifying Person and Number

As discussed in section 4.1.1, person and number information are not realized in the same, or even adjacent, slots in Odawa. The separation of person and number information is most extreme in transitive verbs with animate objects, where person and number of both subject and object are discontinuous. This can be seen in (3a), reproduced here with first person/number affixes bolded and third person/number affixes underlined, as in ***n**-waabm-<u>aa</u>-swaa-**mnaa**-dgen-<u>ag</u>* 'perhaps we don't see them'.

The separation of person and number can be inconvenient for non-specialist use of the analyzer, since it is customary to refer to person-number combinations as atomic entities (e.g. first person plural form), or impractical in its integration with other software applications, which may need to know only the set of morphological features, in some standard form and order, expressed by a word-form, instead of its exact morphological break-down. To address this, we have produced a second version of our model that translates the low-level analyses from the core model into a form with a standardized set of morphological features presented in a canonical order.

5.4 Model Behavior Examples

To summarize, in effect we have created two models, a basic one that provides a what-you-see-is-what-you-get parse of the morphology, and another that interprets the basic parse into a more condensed form. Both versions of our model carry out the full set of phonological mappings, including the vowel deletion process mentioned above. Concretely, this means that our models return the indicated analyses for the examples in (9), where the first analysis is the basic analysis and the second is the abstracted one.[7]

(9) a. bgizo
swim-3
swim-3.SG
'He swims'

b. bgiz-wi-bn-iig
swim-3-PRF-3.PL
swim-PRF-3.PL
'They have swum'

c. n-bagiz
1-swim
swim-1.SG
'I swim'

d. n-bagzo-mnaa-ba
1-swim-1.PL-PRF
swim-PRF-1.PL
'We (excl.) have swum'

[7] Strictly speaking, our models return a lemma, rather than an English translation. Also, the full translated analyses include overt specification of default, unmarked features, like `+Pos` for verbs with positive polarity. These are suppressed here for brevity

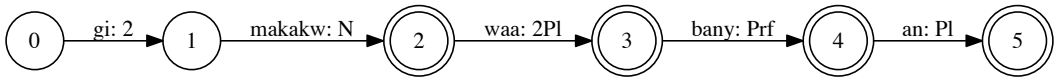

Figure 2: Finite state machine corresponding to path in Odawa model that recognizes *gmakkowaabniin* 'your folks' former boxes'. Further phonological processing allows the morpheme sequence in the model to match the actually attested form.

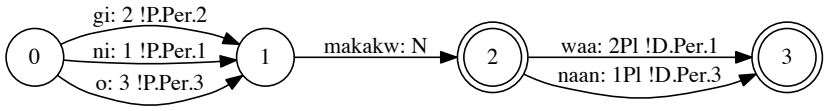

Figure 3: Finite state machine with memory to enforce co-occurrence restrictions between prefixes and suffixes.

As can be seen in (9), subject information appears string-finally in parses produced by the more abstract model. The only exception to this is in transitive verbs, which inflect for person and number of both subjects and objects. In this case, object information occurs string-finally and is marked with an explicit object marker. This is exemplified in (10), which reproduces (3a).

(10) n-waabm-aa-swaa-mnaa-dgen-ag
1-see-3.OBJ-NEG-1.PL-DUB-3.PL
see-NEG-DUB-1.PL-3.PL.OBJ
'Perhaps we (excl.) don't see them'

Note also that Algonquian languages are characterized by an imperfect correspondence between morphological slots and subject-object agreement. That is, while (10) shows agreement with a first person plural exclusive subject with *ni-...-mnaa* '1-...-1.PL', these same morphemes index object features in (11).

(11) n-waabm-ig-sii-mnaa-dgen-ag
1-see-3.SUB.PRE=OBJ-NEG-1.PL-DUB-3.PL
see-NEG-DUB-3.PL-1.PL.OBJ
'Perhaps they don't see us (excl.)'

Our translation module manages the distinction between (10) and (11) by identifying which morpheme occurs in the post-root slot (called a 'theme-sign' in the Algonquianist tradition). Hence, in our examples, if the suffix is *-aa*, then the object is third person and the subject is indexed by the prefix, while if it is */-igo/*, the subject is third person and the object is indexed by the prefix.[8] Such an approach allows our model to avoid global computations of the alignment of person hierarchies and grammatical role hierarchies commonly discussed in the descriptive literature (e.g. Valentine 2001:267-272).

5.5 Performance

We tested our core model against a small corpus (7,578 tokens, 2,685 types) of written Odawa. Our corpus consists of narratives and sentences collected in 1937 by Leonard Bloomfield (1957), less 45 unassimilated English loan types. This collection of texts is a somewhat atypical testing ground for language model development. First, Valentine (2001), the description which our model is based on, draws heavily from the collection, making it

[8] For the curious reader, the unabridged output for (10) appears below for the basic and translated versions, respectively:

```
1+waabmaa+VTA+ThmDir+Neg+1+Pl+Dub+3+Pl
waabmaa+VTA+Ind+Neg+Dub+1Pl+3PlO
```

For (11) the model outputs are:

```
1+waabmaa+VTA+ThmInv+Neg+1+Pl+Dub+3+Pl
waabmaa+VTA+Ind+Neg+Dub+3Pl+1PlO
```

In both versions '+VTA' is a common Algonquianist abbreviation for transitive verb with an animate object, in the basic model '+ThmDir' and '+ThmInv' are notations for third person object agreement and third person subject agreement, respectively (abbreviating Algonquianist 'direct theme' and 'inverse theme'). In the translated tags '+Ind' signals that the verb is a matrix clause form (Algonquianist 'independent order').

impossible to truly withhold data for use in testing. Furthermore, though Valentine (2001) draws many examples from the text collection, Andrew Medler, Bloomfield's consultant, spoke a dialect that differs slightly from the phonology of modern Odawa described in Valentine (2001). Finally, while the spelling system in the texts has been updated, no attempt was made to correct subphonemic effects in Bloomfield's transcription.

With small changes to compensate for deviations of Medler's speech from the modern dialects described by Valentine (2001), the core model recognized 84% of types in our corpus. To date, our corpus has no hand-verified analyses, making it impossible to provide precision or recall statistics.

A preliminary survey of errors indicated that the errors are distributed as shown in (12). Errors were identified as 'lexical' if a lexical item needed to be added or modified for successful parsing, 'morphological' if the error was a result of a missing morphological process, 'phonological' if a phonological rule had misapplied or was absent, 'orthographical' if the error resulted from an orthographic convention (these were overwhelmingly the misuse of apostrophes or the omission of interconsonantal *n*), 'dialectal' if the error resulted from dialect differences, and 'other' if the above categories did not apply.

(12)

Type	% of Errors
orthography	30%
morphology	28%
lexicon	20%
phonology	13%
dialect	5%
other	2%

Our core model parsed the full 7,578 word corpus in an average of 0.653 seconds over 15 runs on an Intel core i7@2.9 GHz processor, which equals a parsing speed of 11,605 words per second.

6 Applications

Morphological parsers, while useful for linguists, enable the creation of many downstream applications that have usefulness for a much broader audience. In our experience, one of the greatest benefits is found by being able to augment an electronic dictionary with the parser, creating an "intelligent" dictionary (Johnson et al. 2013). It has long been noted that, for languages with rich inflectional morphology, a morphologically non-aware dictionary can be extremely cumbersome to use, especially for speakers, learners and others lacking many years of linguistic training.

With a morphological parser, however, users may input any inflected form of a word, and be redirected to the appropriate lemma (cf. `http://www.nishnaabemwin.atlas-ling.ca`). While the user still may need to grapple with understanding some amount of linguistic terminology in order to fully benefit from the parse (e.g. identifying a particular form as the '2nd person plural dubitative negative' is still less than completely helpful for many users), at least they will be directed to the correct lexical entry, and so will be able to retrieve the appropriate lexical content of the word, even if its grammatical specifications are still somewhat opaque.

Moreover, even the grammatical information in the parse can itself be presented in a more user-friendly form. The same '2nd person plural dubitative negative' form could equally well be presented as the 'you folks perhaps do not X' form. Thus, although the inner workings of the electronic dictionary and parser remain highly technical, the view presented to the user can be made much more welcoming.

Furthermore, the morphological parser can also be used in reverse to generate individual word-forms expressing some desired combination of morphological features. Naturally, this can be scaled up to present collections of inflected forms (e.g. core word-forms or the full inflectional paradigm).

Morphological parsers can also facilitate the use and production of texts (cf. `http://altlab.ualberta.ca/korp/`). In the Odawa communities we work with, there is high demand from students for lemmatization of texts. The linking of an inflected word to the lemma in an electronic dictionary uses the same mechanism as lemmatization of texts, making this operation straightforward (`http://altlab.ualberta.ca/kidwinan/`). If the text is in an electronic format, the parser can even provide an on-the-fly analysis of the morphology of a word.

The benefits of an application of this sort are manifold. In particular, it allows learners (and newly-literate speakers) the chance to explore a wide range of texts, challenging their reading abil-

ities and discovering new words on their own. This is especially valuable in languages which lack graded readers, and where people may be motivated to engage with texts that are above their proficiency levels in order to extract culturally-relevant information contained within. While the on-the-fly lemmatization is no substitute for a fully-annotated interlinear gloss, it is still a powerful aid in the development of written comprehension, which itself may increase the demand for more and better texts to be produced in the language.

Furthermore, parsers define a set of legal words, and therefore underlie important tools like spell-checkers and grammar checkers. Such tools can be helpful for literacy programs and speed the creation and proofing of high-quality texts in the language. Where communities are attempting to promulgate a particular written form of the language as standard, such tools can help in the codification and enforcement of those standards.

It is a short leap from the applications described above to classroom applications as well. Foremost among these are intelligent computer-aided language learning (or I-CALL) applications (Antonsen et al. 2013). The combination of a lexicon, a morphological parser and some simple grammatical rules can allow for the creation of an essentially infinite number of language drills of various types.

Because of the morphological knowledge that the parser contains, it is possible to give students feedback on their responses that goes well beyond "right" and "wrong". For example, an I-CALL application can recognize that although the drill is calling for the first person plural form of the verb to be provided, the student has instead offered the second person singular form. The application could then provide that feedback to the student, letting them know that although the form they gave was incorrect, it was in fact a valid form in the language.

In the longer term, the application can keep track of the students' responses, allowing the developers to analyze the patterns of correct and incorrect answers. This provides invaluable information for curriculum developers as they field-test new courses. This is especially important for developers working with endangered languages, where there is typically little to no pedagogical tradition to follow. Being able to apply quantitative measures to questions such as "Is it better to teach declarative forms before imperative forms, or the other way around?" has great potential for improving the efficacy of language teaching programs. Given the vital role that such programs play in the long-term resurgence of endangered languages, the potential benefits of these applications should not be discounted.

7 Conclusion

The ALT-lab group at the University of Alberta is developing language technology for First Nations languages. Our most recent project is a morphological parser of the Odawa dialect of Ojibwe, which is currently in an advanced beta stage. This parser comes in two versions, one which closely follows the morphology of the language, and another which interprets and reorganizes the morphology into a more user-friendly format. The development of a parser opens the door to exciting new research and opportunities for community applications.

References

Antonsen, L., T. Trosterud, and H. Uibo (2013). Generating modular grammar exercises with finite-state transducers. In *Proceedings of the second workshop on NLP for computer-assisted language learning at NODALIDA 2013. NEALT Proceedings Series 17 / Linkping Electronic Conference Proceedings 86*, pp. 2738.

Arppe, A., C. Harvey, M.-O. Junker, and J. R. Valentine (fc.). Algonquian verb paradigms. a case for systematicity and consistency. In *Algonquian Conference 47*.

Baraga, F. (1878a). *A Dictionary of the Otchipwe Language: Explained in English* (Second ed.). Beauchemin and Valois.

Baraga, F. (1878b). *A Theoretical and Practical Grammar of the Otchipwe Language* (Second ed.). Beauchemin and Valois.

Beesley, K. R. and L. Karttunen (2003). *Finite State Morphology*. CSLI Publications.

Bloomfield, L. (1957). *Eastern Ojibwa: Grammatical Sketch, Texts and Word List*. Ann Arbor: University of Michigan Press.

Bowers, D. (2015). *A System for Morphophonological Learning and its Consequences for Language Change*. Ph. D. thesis, UCLA.

Brittain, J. (1999). A reanalysis of transitive animate theme signs as object agreement: Evidence from western naskapi. In *Papers of the 30th Algonquian Conference*.

Chomsky, N. and M. Halle (1968). *The Sound Pattern of English*. Harper and Row.

Corbiere, M. A. and J. R. Valentine (2016). Nishnaabemwin: Odawa and Eastern Ojibwe online dictionary.

Cuoq, J. A. (1886). *Lexique de la langue Algonquine*. J Chapleau et fils.

Golla, V. (2007). North America. In *Encyclopedia of the World's Endangered Languages*. Routledge.

Huldén, M. (2009). Foma: A finite state toolkit and library. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 29–32.

Johnson, C. D. (1972). *Formal Aspects of Phonological Description*. Mouton.

Johnson, R., L. Antonsen, and T. Trosterud (2013). Using finite state transducers for making efficient reading comprehension dictionaries. In S. Oepen, K. Hagen, and J. B. Johannessen (Eds.), *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013). NEALT Proceedings Series 16*, pp. 59–71.

Jurafsky, D. and J. Martin (2000). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice-Hall.

Kaplan, R. M. and M. Kay (1994). Regular models of phonological rule systems. *Computational Linguistics 20*, 331–378.

Koskenniemi, K. (1983). *Two-Level Morphology: A General Computational Model for Word-Form Recognition and Production*. University of Helsinki.

Lindén, K., E. Axelson, S. Hardwick, M. Silfverberg, and T. Pirinen (2011). Hfst: Framework for compiling and applying morphologies. In *Proceedings of the Second International Workshop on Systems and Frameworks for Computational Morphology*, pp. 67–85.

McGinnis, M. (1999). Is there syntactic inversion in Ojibwa? In L. Bar-el, R. Dechaine, and C. Reinholtz (Eds.), *Papers from the Workshop on Structure & Constituency in Native American Languages*, Volume MIT Occasional Papers in Linguistics 17, pp. 101–118.

Nichols, J. D. (1980). *Ojibwe Morphology*. Ph. D. thesis, Harvard.

Piggott, G. L. and J. Kaye (1973). Odawa language project: Second report. Technical report, University of Toronto.

Piggott, G. L., J. Kaye, and K. Tokaichi (1971). Odawa language project: First report. Technical report, University of Toronto.

Rhodes, R. (1985a). *Eastern Ojibwa-Chippewa-Ottawa Dictionary*. Mouton.

Rhodes, R. (1985b). Lexicography and Ojibwa Vowel Deletion. *The Canadian Journal of Linguistics 30*(4), 453–471.

Snoek, C., D. Thunder, K. Lõo, A. Arppe, J. Lachler, S. Moshagen, and T. Trosterud (2014). Modeling the noun morphology of Plains Cree. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*.

Trosterud, T. (2005). Grammatically based language technology for minority languages. In *Lesser-Known Languages of South Asia: Status and Policies, Case Studies and Applications of Information Technology*, pp. 293–316. Mouton de Gruyter.

Valentine, J. R. (2001). *Nishnaabemwin Reference Grammar*. Toronto: University of Toronto Press, Inc.