Modeling Northern Haida Verb Morphology

Jordan Lachler[1], Lene Antonsen[2], Trond Trosterud[2], Sjur N. Moshagen[2], Antti Arppe[1]
[1]University of Alberta
[2]UIT Arctic Univerity of Norway
Email: {lachler, arppe}@ualberta.ca
{lene.antonsen, trond.trosterud, sjur.n.moshagen}@uit.no

Abstract

This paper describes the development of a computational model of the morphology of Northern Haida based on finite state machines (FSMs), with a focus on verbs. Northern Haida is highly endangered, and a member of the isolate Haida macrolanguage, spoken in British Columbia and Alaska. Northern Haida is a highly-inflecting language whose verbal morphology relies largely on suffixes, with a limited number of prefixes. The suffixes trigger morphophonological changes in the stem, participate in blocking, and exhibit variable ordering in certain constructions. The computational model of Northern Haida verb morphology is capable of handling these complex affixation patterns and the morphophonological alternations that they engender. In this paper, we describe the challenges we encountered and the solutions we propose, while contextualizing the endeavour in the description, documentation and revitalization of First Nations Languages in Canada.

Keywords: less-resourced/endangered languages, language modeling, morphology

1. Introduction

While the study of the Indigenous languages of North America has been a mainstay of descriptive linguistics over the past century or more, these languages have received much less attention in the computational linguistic literature. This has begun to change in recent years, e.g. through work initiated on Algonquian languages such as Plains Cree (Snoek et. al., 2014; Harrigan et. al., 2017), East Cree (Arppe et al., 2017a), Odawa (Bowers et al. 2017), as well as Dene languages such as Tsuut'ina (Arppe et al. 2017b). This work has come about through computational linguists, field linguists and Indigenous communities working together to create tools and applications, which can support the continued use and learning of these often threatened languages (Arppe et al., 2016). In this paper, we provide a description of a computational model of the verbal morphology of one highly endangered North American Indigenous language, Northern Haida, which was created in furtherance of these goals.

2. Background

2.1. Northern Haida

Northern Haida, known as *Xaad Kíl*, is one of two members of the macrolanguage Haida (Simons and Fennig, 2017). Speakers of Northern Haida live in a small number of communities on the Haida Gwaii archipelago off the coast of British Columbia, and on the islands of far southeastern Alaska. Haida is an isolate, with no known linguistic relatives. Today, there are fewer than a dozen fluent speakers of Northern Haida remaining, all of them over the age of 70, but also an active and growing cohort of younger learners of the language.

2.2. Morphological Structure of Verbs

Most verbs in Northern Haida are entirely suffixing in a relatively neatly agglutinative manner. They inflect for a range of grammatical categories including Tense, Aspect, Mood, Evidentiality, and Polarity among others, but not for the person and number values of their subject or object. The only exception to this is the Third Person Plural suffix *-'wa-*, which is used to indicate that there is a definite third person plural participant in the clause.

In total, there are fewer than 20 distinct suffixes, but they can combine to give nearly 200 basic paradigmatic forms for each verb. While the pattern of suffixes is quite agglutinative in some respects, there are several complications which are worthy of note.

To begin with, the suffixes frequently trigger morphophonemic changes in the verb stem to which they attach (and vice versa). For instance, the Present tense suffix has three forms, *-ng*, *-ang* and *-gang*. The form *-ang* occurs only after verb stems that end in either *-s* or *-d*. The addition of the Present suffix to verb stems of this type causes the last vowel of the stem to change to *-ii-*, as with the verb *k'agáangad* "oversleep" and its Present tense form *k'agáangiidang*.

Moreover, while the order of suffixes is not variable, it is also not entirely consistent. For example, the Negative suffix {*-'ang-*} and the Habitual aspect suffix {*-gang-*} occur in different orders, depending on what other suffixes are present. In the Present tense, the Habitual comes first, followed by the Negative: *gatáa-gang-'ang-gang* (eat-HAB-NEG-PRES) "never eats". However, in the Reported Past, the order is reversed,

with the Negative coming first and the Habitual coming second: *gatáa-'ang-gaang-aa-n* (eat-NEG-HAB-REP-PAST) "never used to eat".

A further complication arises from certain logical suffix combinations which are blocked. The Habitual aspect suffix and the Future tense suffix do not occur together, even though they are not semantically incompatible. The Habitual aspect suffix also does not occur adjacent to the Past tense suffix. Instead, the single Customary tense-aspect suffix {*-giinii*} is used: *gatáa-giinii* (eat-CUST) "used to eat". Note that this only applies when those two suffixes would be adjacent to one another; when other suffixes intervene (as in the example of *gatáa'anggaangaan* above), then both suffixes can and do occur.

While the large majority of Northern Haida verbs are solely suffixing, as described above, a significant minority have an obligatory prefix slot as well. This slot is filled by a Classifier prefix, which gives further semantic specification about the absolutive argument in the clause. For instance, the verb stem CLSF+*gang* "hold something of CLSF-shape" require a monosyllabic prefix which indicates what type of object is behind held. Typical forms include: *dlagáng* "hold something animate", *tlagáng* "hold something thin and flat", *skáagang* "hold something small and round", *k'íigang* "hold something large", and *chagáng* "hold a bag of something". The number of possible Classifiers varies by verb, from just a few, to a couple dozen, to several hundred. These verbs are then suffixally inflected in the same way as other verbs.

A final complication in the conjugation of Northern Haida verbs is its robust system of Auxiliaries. Auxiliaries occur immediately following the main verb. The main verb appears in one of several Construct states, depending on the identity of the following Auxiliary. The inflections, then, manifest as suffixes on the Auxiliary. For example, the main verb *k'ajúu* "sing" can be followed the Auxiliary *áwyaa* "very, really". When these combine, *k'ajúu* occurs in its Construct form *k'ajáaw*, and the inflectional suffixes occur on the Auxiliary *áwyaa*: *k'ajáaw áwyaa-gan* (sing:CONST very-PAST) "really sang". There are around 40 of these Auxiliaries, which can sometimes stack up after the main verb. The result is a multi-part verb form, written orthographically as two or more words, but bearing a single inflection, e.g.: *káayd déed gudáa-ng* (leave right.away want.to-PRES) "wants to leave right away". There are around 160 attested Auxiliary strings, each of which can in principle be inflected for the full range of nearly 200 endings found on main verbs. Thus, including the possible Auxiliaries, each verb phrase can in principle be inflected for approximately 32,000 distinct forms.

3. Computational Modeling

3.1. Framework

As the computational formalism for implementing our model for Northern Haida morphology, we make use of Finite-state machines (FSMs) (cf. e.g. Beesley & Karttunen, 2003) which have become one standard way for computationally modeling the morphological structure of words in natural languages. There are currently several implementations of FSM compilers, e.g. *xfst* (Beesley & Karttunen 2003), *foma* (Hulden 2009) and HFST (Lindén et al. 2011), of which the first is available for non-commercial research use and the second and third are open source resources. The key advantages of FSMs are many, but most crucially they are designed for rule-based definition of paradigms, which does not require large corpora from which to learn such rules. This is helpful in the case of endangered languages, such as Northern Haida, for which such corpora are usually lacking.

Furthermore, FSMs allow for easy integration with other software applications, for instance as spell-checking modules within word-processors, morphologically “intelligent” electronic dictionaries, and “intelligent” computer-aided language-learning applications. Here, we make use of the *Giella* infrastructure, developed by the Giellatekno and Divvun research teams at the University of Tromsø (Trosterud 2006; Moshagen et al. 2013), which provides ready-made solutions for the integration of an FSM-based computational model as part of such end-user applications.

3.2. Design considerations and choices

In designing a finite-state computational model, one has to decide whether to model morphophonological alternations at stem+affix junctures by (1) dividing stems into subtypes which are each associated with their own inflectional affix sets that can simply be glued onto the stem, or by (2) modeling such morphophonological alternations using context-based rewrite rules. Furthermore, one has to decide the extent to which one treats affix sequences by splitting these into their constituent morphemes, each associated with one morphosyntactic feature, or rather treats affixes as unanalyzed chunks which are associated with multiple morphosyntactic features (Arppe et al., 2018; Arppe et al. 2017a, 2017b). The more one splits affix sequences, the more one may need to develop and test rules for dealing with morphophonological alternations at these morpheme junctures, whereas in the case of chunking such alternations are precomposed within the chunk. In contrast, the more one uses chunks, the more one has to enumerate chunks based on the number of relevant inflectional subtypes.

While the chunking strategy is not parsimonious and compact in terms of linguistic description, in our experience it results in FST source code which is

nevertheless structurally quite flat and easily comprehensible for scholars who are not specialists in the language in question. Importantly, current finite-state compilers, e.g. *xfst*, HFST, or *foma* (Beesley and Karttunen 2003; Lindén et al. 2011; Hulden 2009), implement a minimization procedure on the finite-state model, so that recurring realizations of string-final character sequences and associated morphological features are systematically identified and merged, resulting, in the end, in a relatively compact model. On the other hand, if some aspect of the chunked morpheme sequences needs to be changed, with the chunking strategy these have to be implemented in multiple locations.

For the Northern Haida model, we decided to (1) split the pre-stem morphemes (the classifiers), as there are very few morphophonological phenomena and these are very regular; (2) entirely chunk the post-stem suffix morphemes, associating the chunks with multiple morphological feature tags; and (3) make maximal use of inflectional subtypes using technical stems and post-stem technical suffix chunks, an approach which we call *maximal chunking*. Thus, we will require only a few morphophonological rules for the stem-suffix morpheme juncture. These morphophonological rules are implemented using the TWOLC formalism within the FST framework. As to the rest, the LEXC formalism in the FST framework is used to define the concatenation of stems and affixal morphophology (with morpheme sequences treated chunks). There are currently just under 10,000 verb stems in the lexicon (Enrico, 2005; Lachler, 2010).

We now discuss the motivations for these design choices (which are similar to, and indeed the original inspiration for those adopted for East Cree, cf. Arppe et al. 2017b) from the perspective of documentary linguists working on an endangered language, and their practical implementation. The morphophonemic changes that the endings trigger on the verb stems are limited to the rhyme of the final syllable of the stem. As such, verb stems are grouped into sub-lexica by the shape of their final rhyme. They are listed with all but their final rhymes in the lexicon, which we term the *technical stem*. The technical stem is then augmented with various final rhymes to produce an inflectable stem.

For example, many verbs have a lexical stem that ends in a short *-a-* (e.g. *skyáana* "be awake"). Before certain endings, the final *-a-* remains short, while before other endings it lengthens to *-aa-*. As such, these verbs are listed in the lexicon with their technical stem which does not include this variable-length final vowel.

```
skyáan CLASS-A "be awake" ;
```

The technical stem then directs to the appropriate continuation lexicon (in this case, CLASS-A, named after the form of the underlying rhyme). Here, two stems are created, one with a short *-a-* and another with a long *-aa-*.

```
LEXICON CLASS-A
 :a CLASS-A-STEM-1 ;
 :aa CLASS-A-STEM-2 ;
```

These two stems then each direct to different continuation lexica, one containing all of the endings which condition a final short *-a-* in the stem (such as in the Present tense form *skyáanang*), and the other containing all of the endings which condition a final long *-aa-* in the stem (such as in the Future tense form *skyáanaasaang*).

```
LEXICON CLASS-A-STEM-1
+V+PRES:ng # ;
LEXICON CLASS-A-STEM-2
+V+FUT:saang # ;
```

We next illustrate how our maximal chunking model works in practice. For example, the inflected form *skyáana'ang'ugan* "were not awake" would be composed from the CLASS-A-STEM-1 form *skyáana* and the following ending or "chunk":

```
LEXICON CLASS-A-STEM-1
+V+NEG+3PL+DIR+PAST:'áng'ugan # ;
```

This ending carries the grammatical features of negative polarity (+NEG), a third person plural participant in the clause (+3PL), direct evidentiary knowledge (+DIR) and past tense (+PAST).

The addition of this ending yields the form *skyáana'áng'ugan*, with two accented syllables, which is incorrect. This triggers a TWOLC rule which deletes any accent after the first accent in the word, producing the correct form *skyáana'ang'ugan*.

While it is straightforward to map those four grammatical tags onto the three component suffixes of that ending (*-'ang-* Negative, *-'u-* 3rd Person Plural, and *-gan* Direct Past), we took the maximal chunking approach in order to minimize the development time required to create the morphological model. Although all of the complications described above in Section 2.2 can be easily handled within LEXC and TWOLC using a non-chunking approach, significant time is required for the field linguist to learn all of the formalisms required to handle morphophonemic alternations, morpheme blocking, and any context-dependent morpheme orderings. By adopting a chunking approach, the morphological model more closely mirrors the field linguist's original documentation of the language. This, in turn, streamlines the development and testing of the FSM, allowing the team to move more rapidly towards the creation of practical tools and applications.

4. Evaluation

For evaluation, we created complete inflectional paradigms for each of the inflectional subtypes, consisting of pairing of the realized surface forms and the underlying morphosyntactic analyses. Due to our maximal chunking approach and the comprehensiveness of the lexical database we have at our disposal, we have been able to reach 100% accuracy for both word-form analysis and generation.

5. Conclusion

The development of computational models for morphologically complex endangered languages can often be slowed due to the difficulty of adapting existing language descriptions into the appropriate computational formalisms. The approach offered here for Northern Haida seeks to minimize any such delay by simplifying the formal machinery required to produce correct results, even where this may run afoul of descriptive parsimony. This should allow more descriptive linguists to develop computational models of the language they have documented, and to move more rapidly to the stage where useful tools and applications can be deployed in the community.

6. Acknowledgements

This work has been undertaken in the Alberta Language Technology Lab (ALTLAb: URL: http://altlab.artsrn.ualberta.ca), University of Alberta, as part of a more general research program to computationally model North American Indigenous languages (Arppe et al. 2016), and it has been funded by the Social Sciences and Humanities Research Council of Canada (SSHRC) through a Partnership Development Grant (890-2013-0047), a Connections Outreach Grant (611-2016-0207), and a KIAS Research Cluster Grant 2015-2018 (University of Alberta).

Bibliographical References

Arppe, A., Cox, C., Hulden, M., Lachler, J., Moshagen, S. N., Silfverberg, M. and Trosterud, T. (2017). Computational Modeling of Verbs in Dene Languages: The Case of Tsuut'ina. *Proceedings of the 2016 Dene Languages Conference*, "Red Books" series, Alaska Native Language Center, University of Alaska, Fairbanks.

Arppe, A., Junker, M.-O. Harvey, C., and Valentine, J. R. (2018). Algonquian verb paradigms. a case for systematicity and consistency. *Papers of the Algonquian Conference 47 (PAC47),* 1-22.

Arppe, A., Junker, M-O, and 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 Computational Methods for Endangered Languages (ComputEL-2)*, 43-47, University of Hawai'i, Manoa, 6-7 March 2017. ACL Anthology.

Arppe, A., Lachler, J., Trosterud, T., Antonsen, L., and Moshagen, S. N. (2016). Basic Language Resource Kits for Endangered Languages: A Case Study of Plains Cree. In: Soria, C., Pretorius, L., Declerck, T. Mariani, J., Scannell, K., and Wandl-Vogt, E. (eds). Proceedings of the LREC 2016 Workshop *CCURL 2016 – Collaboration and Computing for Under-Resourced Languages – Towards an Alliance for Digital Language Diversity*, (Portorož, Slovenia, 23 May 2016), 1-8.

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

Bowers, D., Arppe, A., Lachler, J., Moshagen, S. N. and Trosterud, T. (2017). A Morphological Parser for Odawa. *Proceedings of the 2nd Workshop on Computational Methods for Endangered Languages (ComputEL-2)*, 1-9, University of Hawai'i, Manoa, 6-7 March 2017. ACL Anthology.

Enrico, J. (2005). *Haida Dictionary*. University of Alaska-Fairbanks and Sealaska Heritage Institute.

Harrigan, A., Schmirler, K., Arppe, A., Antonsen, L. Moshagen, S. N., Trosterud, T. and Wolvengrey, A. (accepted pending revisions 22-Jan-2017; revised version submitted 29-Sept-2017). Learning from the Computational Modeling of Plains Cree Verbs. *Morphology,* 27(4), 565–598.

Hulden, M. (2009). Foma: a finite-state compiler and library. In *Proceedings of EACL*, pp. 29–32, Athens. Association for Computational Linguistics.

Lachler, J. (2010). *Alaskan Haida Dictionary*. Sealaska Heritage Institute.

Lindén, K., E. Axelson, S. Hardwick, M. Silfverberg, and Pirinen, T. (2011). HFST – Framework for Compiling and Applying Morphologies. *Proceedings of Second International Workshop on Systems and Frameworks for Computational Morphology (SFCM)*, 67–85.

Moshagen, S. N., Pirinen, T. and Trosterud, T. (2013). Building an open-source development infrastructure for language technology projects. *Proceedings of the 19th Nordic Conference of Computational Linguistics* (NODALIDA 2013). Linköping Electronic Conference Proceedings #85: 343-352.

Simons, Gary F. and Charles D. Fennig (eds.) (2017). *Ethnologue: Languages of the World*, Twentieth edition. Dallas, Texas: SIL International. Online version: http://www.ethnologue.com.

Snoek, C., Thunder, D., Lõo, K., Arppe, A., Lachler, J., Moshagen, S., and Trosterud, T. (2014). Modeling the Noun Morphology of Plains Cree. In *Proceedings of ComputEL: Workshop on the use of computational methods in the study of endangered languages*, 52nd Annual Meeting of the Association for Computational Linguistics, Baltimore, Maryland, 26 June 2014, 34-42.

Trosterud, T. (2006). Grammatically based language technology for minority languages. In *Lesser known languages of South Asia*, 293–316. Mouton de Gruyter, The Hague.