An FST morphological analyzer for the Gitksan language

Clarissa Forbes[α] **Garrett Nicolai**[β] **Miikka Silfverberg**[β]
[α]University of Arizona [β]University of British Columbia
`forbesc@email.arizona.edu` `first.last@ubc.ca`

Abstract

This paper presents a finite-state morphological analyzer for the Gitksan language. The analyzer draws from a 1250-token Eastern dialect wordlist. It is based on finite-state technology and additionally includes two extensions which can provide analyses for out-of-vocabulary words: rules for generating predictable dialect variants, and a neural guesser component. The pre-neural analyzer, tested against interlinear-annotated texts from multiple dialects, achieves coverage of (75-81%), and maintains high precision (95-100%). The neural extension improves coverage at the cost of lowered precision.

1 Introduction

Endangered languages of the Americas are typically underdocumented and underresourced. Computational tools like morphological analyzers present the opportunity to speed up ongoing documentation efforts by enabling automatic and semi-automatic data analysis. This paper describes the development of a morphological analyzer for Gitksan, an endangered Indigenous language of Western Canada. The analyzer is capable of providing the base form and morphosyntactic description of inflected word forms: a word *gupdiit* 'they ate' is annotated `gup-TR-3PL`.

Our Gitksan analyzer is based on two core documentary resources: a wordlist spanning approximately 1250 tokens, and an 18,000 token interlinear-annotated text collection. Due to the scarcity of available lexical and corpus resources, we take a rule-based approach to modeling of morphology which is less dependent on large datasets than machine learning methods. Our analyzer is based on finite-state technology (Beesley and Karttunen, 2003) using the *foma* finite-state toolkit (Hulden, 2009b).

Our work has three central goals: (1) We want to build a flexible morphological analyzer to supplement lexical and textual resources in support of language learning. Such an analyzer can support learners in identifying the base-form of inflected words where the morpheme-to-word ratio might be particularly high, in a way not addressed by a traditional dictionary. It may also productively generate inflected forms of words. (2) We want to facilitate ongoing efforts to expand the aforementioned 1250 token wordlist into a broad-coverage dictionary of the Gitksan language. Running our analyzer on Gitksan texts, we can rapidly identify word forms whose base-form has not yet been documented. An analyzer can also help automate the process of identifying sample sentences for dictionary words, the addition of which substantially increases the value of the dictionary. (3) We want to use the model to further our understanding of Gitksan morphology. Unanalyzeable and erroneously analyzed forms can help us identify shortcomings in our description of the morphological system and can thus feed back into the documentation effort of the language.

The Gitksan-speaking community recognizes two dialects: Eastern (Upriver) and Western (Downriver). Our analyzer is based on resources which mainly represent the Eastern dialect. Consequently, our base analyzer achieves higher coverage of 71% for the Eastern dialect as measured on a manually annotated test set. For the Western dialect, coverage is lower at 53%. In order to improve coverage on the Western variety, we explore two extensions to our analyzer. First, we implement a number of dialectal relaxation rules which model the orthographic variation between Eastern and Western dialects. This leads to sizable improvements in coverage for the Western dialect (around 9%-points on types and 6%-points on tokens). Moreover, the precision of our analyzer remains high both for the Eastern and West-

Proceedings of the Seventeenth SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology,pages 188–197
August 5, 2021. ©2

ern dialects even after applying dialect rules. Secondly, we extend our FST morphological analyzer by adding a data-driven neural guesser which further improves coverage both for the Eastern and Western varieties.

2 The Gitksan Language

The Gitxsan are one of the indigenous peoples of British Columbia, Canada. Their traditional territories consist of upwards of 50,000 square kilometers of land along the Skeena River in the BC northern interior. The Gitksan language is the easternmost member of the Tsimshianic family, which spans the entirety of the Skeena and Nass River watersheds to the Pacific Coast. Today, Gitksan is the most vital Tsimshianic language, but is still critically endangered with an estimated 300-850 speakers (Dunlop et al., 2018).

The Tsimshianic family can be broadly understood as a dialect continuum, with each village along these rivers speaking somewhat differently from its neighbors up- or downstream, and the two endpoints being mutually unintelligible. The six Gitxsan villages are commonly divided into two dialects: East/Upriver and West/Downriver. The dialects have some lexical and phonological differences, with the most prominent being a vowel shift. Consider the name of the Skeena River: *Xsan, Ksan* (Eastern) vs *Ksen* (Western).

2.1 Morphological description

The Gitksan language has strict VSO word order and multifunctional, fusional morphology (Rigsby, 1986). It utilizes prefixation, suffixation, and both pro- and en-cliticization. Category derivation and number marking are prefixal, while markers of argument structure, transitivity, and person/number agreement are suffixal.

The Tsimshianic languages have been described as having word-complexity similar to German (Tarpent, 1987). The general structure of a noun or verb stem is presented in the template in Figure 1. A stem consists of minimally a root (typically CVC); an example is monomorphemic *gup* 'eat'. Stems may also include derivational prefixes or transitivity-related suffixes; compare *gupxw* 'be eaten; be edible'.

In sentential context, stems are inflected for features like transitivity and person/number. Our analyzer is concerned primarily with stem-external inflection and cliticization. The structure of stem-external morphology for the most complex word type, a transitive verb, is schematized in the template in Figure 2; an example word with all these slots filled would be *'naag̱asḵ'otsdiitg̱athl* 'apparently they cut.PL open (common noun)'

On the left edge of the stem can appear any number of modifying 'proclitics'. These contribute locative, adjectival, and manner-related information to a noun or verb, often producing semi- or non-compositional idioms in a similar fashion to Germanic particle verbs.[1] It is often unclear whether these proclitics constitute part of the root or stem, or if they are distinct words entirely. The orthographic boundaries on this edge are consequently sometimes fuzzy. Sometimes clear contrasts are presented, as with the sequence *lax̱-yip* 'on-earth': we see compositional *lax̱ yip* 'on the ground' versus lexicalized *lax̱yip* 'land, territory'. However, the boundary between compositional and idiomatic is not always so obvious, as in examples like (1).

(1) a. *saa-'witxw* (away-come, 'come from')
b. *ḵ'ali-aks* (upstream-water, 'upriver')
c. *xsi-ga'a* (out-see, 'choose')
d. *luu-no'o* (in-hole, 'annihilate')

Inflectional morphology largely appears on the right edge of the stem. The main complexity of Gitksan inflection involves homophony and opacity: a similar or identical wordform often has multiple possible analyses. For example, a word like *gubin* transparently involves a stem *gup* 'eat' and a 2SG suffix *-n*, but the intervening vowel *i* might be analyzed as epenthetic, as transitive inflection (TR), or as a specially-induced transitivizing suffix (T), resulting in three possible analyses in (2). Similarly, a word *gupdiit* involves the same stem *gup* 'eat' and a 3PL suffix *-diit*, but this suffix is able to delete preceding transitive suffixes, resulting in four possible analyses as in (3).

(2) *gubin*
a. gup-2SG
b. gup-TR-2SG
c. gup-T-2SG

(3) *gupdiit*
a. gup-3PL
b. gup-TR-3PL
c. gup-T-3PL
d. gup-T-TR-3PL

[1] E.g. **nach***slagen* 'look up' in German.

Derivation–	Proclitics–	Plural–	**Root**	–Argument Structure

Figure 1: Morphological template of a complex nominal or verbal stem

Proclitics–	**Stem**	–Transitive	–Person/Number	=Epistemic	=Next Noun Class

Figure 2: Morphological template of modification, inflection, and cliticization for a transitive verbal predicate

Running speech in Gitksan is additionally rife with clitics, which pose a more complex problem for morphological modeling. First, there are a set of ergative 'flexiclitics', which are able to either procliticize or encliticize onto a subordinator or auxiliary, or stand independently. The same combination of host and clitic might result in sequences like *n=ii* (1SG=and), *ii=n* (and=1SG), or *ii na* (and 1SG) (Stebbins, 2003; Forbes, 2018).

Second, all nouns are introduced with a noun-class clitic that attaches to the preceding word, as illustrated by the VSO sentence in (4). Here, the proper noun clitic =*s* attaches to the verb but is syntactically associated with *Mary*, and the common noun clitic =*hl* attaches to *Mary* but is associated with *g̲ayt* 'hat'.

(4)

Giigwis		Maryhl		g̲ayt.
giikw-i-t	=s	Mary	=hl	g̲ayt
buy-TR-3.II	=PN	Mary	=CN	hat

'Mary bought a hat.'

Any word able to precede a noun phrase is a possible host for one of these clitics (hence their appearance on transitive verbs in Figure 2).

Finally, there are several sentence-final and second-position clitics. whose distribution is based on prosodic rather than strictly categorial properties; these attach on the right edge of subordinators/auxiliaries, predicates, and argument phrases, depending on the structure of the sentence.

A large part of Gitksan's unique morphological complexity therefore arises not in nominal or verbal inflection, but in the flexibility of multiple types of clitics used in connected speech, and the logic of which possible sequences can appear with which wordforms.

2.2 Resources

The Gitksan community orthography was designed and documented in the Hindle and Rigsby (1973) wordlist (H&R). Though it originally reflected only the single dialect of one of the authors (Git-an'maaxs, Eastern), this orthography is in broad use today across the Gitxsan community for all dialects, as well as neighboring Nisg̲a'a, with some variations. Given the relatively short period that this orthography has been in use, orthographic conventions can vary widely across dialects and writers. In producing this initial analyzer, we attempt to mitigate the issue by working with a small number of more-standardized sources: the original H&R and an annotated, multidialectal text collection.

We worked with a digitized version of the H&R wordlist (Mother Tongues Dictionaries, 2020). The original wordlist documents only the Git-an'maaxs Eastern dialect; our version adds a small number of additional dialect variants, and fifteen common verbs and subordinators. In total, the list contains approximately 1250 lexemes and phrases, plus noted variants and plural forms.

The analyzer was informed by descriptive work on both Gitksan and its mutually intelligible neighbor Nisg̲a'a. This work details many aspects of Gitksan inflection, including morphological opacity and the complex interactions of certain suffixes and clitics (Rigsby, 1986; Tarpent, 1987; Hunt, 1993; Davis, 2018; Brown et al., 2020).

A text collection of approximately 18,000 words was also used in the development and evaluation of the analyzer. This collection consists of oral narratives given by three speakers from different villages: Ansbayaxw (Eastern), Gijigyukwhla'a (Western), and Git-anyaaw (Western) (cf. Forbes et al., 2017). It includes multiple genres: personal anecdotes, traditional tales (*ant'imahlasxw*), histories of ownership (*adaawk̲*), recipes, and explanations of cultural practice. The collection is fully annotated in the 'interlinear gloss' format with free translation, exemplified in (5).

(5)

Ii	al'alg̲altg̲athl	get,
ii	CVC-alg̲al-t=g̲at=hl	get
CCNJ	PL-watch-3.II=REPORT=CN	people

'And they stood by and watched,'

The analyzed corpus provides insight into the use of clitics in running speech, and is the dataset against which we test the results of the analyzer.

3 Related Work

While considering different approaches to computational modeling of Gitksan morphology, finite-state morphology arose as a natural choice. At the present time, finite-state methods are quite widely applied for Indigenous languages of the Americas. Chen and Schwartz (2018) present a morphological analyzer for St. Lawrence Island / Central Siberian Yupik for aid in language preservation and revitalization work. Strunk (2020) present another analyzer for Central Alaskan Yupik. Snoek et al. (2014) present a morphological analyzer for Plains Cree nouns and Harrigan et al. (2017) present one for Plains Cree verbs. Littell (2018) build a finite-state analyzer for Kwak'wala. All of the above are languages which present similar challenges to the ones encountered in the case of Gitksan: word forms consisting of a large number of morphemes, both prefixing and suffixing morphology and morphophonological alternations. Finite-state morphology is well-suited for dealing with these challenges. It is noteworthy that similarly to Gitksan, a number of the aforementioned languages are also undergoing active documentation efforts.

While we present the first morphological analyzer for Gitksan which is capable of productive inflection, this is not the first electronic lexical resource for the Gitksan language. Littell et al. (2017) present an electronic dictionary interface Waldayu for endangered languages and apply it to Gitksan. The model is capable of performing fuzzy dictionary search which is an important extension in the presence of orthographic variation which widely occurs in Gitksan. While this represents an important development for computational lexicography for Gitksan, the method cannot model productive inflection which is important particularly for language learners who might not be able to easily deduce the base-form of an inflected word (Hunt et al., 2019). As mentioned earlier, our model can analyze inflected forms of lexemes.

We extend the coverage of our finite-state analyzers by incorporating a neural morphological guesser which can be used to analyze word forms which are rejected by the finite-state analyzer. Similar mechanisms have been explored for other American Indigenous languages. Micher (2017) use segmental recurrent neural networks (Kong et al., 2015) to augment a finite-state morphological analyzer for Inuktitut.[2] These jointly segment the input word into morphemes and label each morpheme with one or more grammatical tags. Very silmilarly to the approach that we adopt, Schwartz et al. (2019) and Moeller et al. (2018) use attentional LSTM encoder-decoder models to augment morphological analyzers for extending morphological analyzers for St. Lawrence Island / Central Siberian Yupik and Arapaho, respectively.

4 The Model

Our morphological analyzer was designed with several considerations in mind. First, given the small amount of data at our disposal, we chose to construct a rule-based finite state transducer, built from a predefined lexicon and morphological description. The dependence of this type of analyzer on a lexicon supports one of the major goals of this project: lexical discovery from texts. Words which cannot be analyzed will likely be novel lemmas that have yet to be documented. Furthermore, the process of constructing a morphological description allows for the refinement of our understanding of Gitksan morphology and orthographic standards. For example, there is a common post-stem rounding effect that generates variants such as *jog̲at, jog̲ot* 'those who live'; the project helps us identify where this effect occurs. Our analyzer can also later serve as a tool to explore of the behavior of less-documented constructions (e.g. distributive, partitive), as grammatical and pedagogical resources continue to be developed.

Our general philosophy was to take a maximal-segmentation approach to inflection and cliticization: morphemes were added individually, and interactions between morphemes (e.g. deletion) were derived through transformational rules based on morphological and phonological context. Most interactions of this kind are strictly local; there are few long-distance dependencies between morphemes. The only exception to the minimal chunking rule is a specific interaction between noun-class clitics and verbal agreement: when these clitics append to verbal agreement suffixes, they either agglutinate with (6-a) or delete them (6-b) depending on whether the agreement and noun-class morpheme are associated with the same noun (Tarpent, 1987; Davis, 2018). That is, the conditioning factor for this alternation is syntactic, not morphophonological.

[2] The Uquailaut morphological analyzer: http://www.inuktitutcomputing.ca/Uqailaut

(6) Realizations of *gup-i-**t**=hl* (eat-TR-3=CN)

a. *gubi**t**hl* ‘he/she ate (common noun)’

b. *gubihl* ‘(common noun) ate’

The available set of resources further constrained our options for the analyzer’s design and our means of evaluating it. The H&R wordlist is quite small, and of only a single dialect, while the corpus for testing was multidialectal. We therefore aimed to produce a flexible analyzer able to recognize orthographic variation, to maximize the value of its small lexicon.

4.1 FST implementation

Our finite-state analyzer was written in *lexc* and *xfst* format and compiled using *foma* (Hulden, 2009b). Finite-state analyzers like this one are constructed from a dictionary of stems, with affixes added left-to-right, and morpho-phonological rewrite rules applied to produce allomorphs and contextual variation. The necessary components of the analyzer are therefore a lexicon, a morphotactic description, and a set of morphophonological transformations, as illustrated in Figure 3.

Our analyzer’s lexicon is drawn from the H&R wordlist. As a first step, each stem from that list was assigned a lexical category to determine its inflectional possibilities. The resulting 1506 word + category pairs were imported to category-specific groups in the morphotactic description.

Any of the major stem categories could be used to start a word; modifiers, preverbs, and prenouns could also be used as verb/noun prefixes. Each categorized group flowed to a series of category-specific sections which appended the appropriate part of speech, and then listed various derivational or inflectional affixes that could be appended. A morphological group would terminate either with a hard stop (#) or by flowing to a final group ‘Word’, where clitics were appended.

Finally, forms were subject to a sequence of orthographic transformations reflecting morphophonological rules. Some examples included the deletion of adjacent morphemes which could not co-occur, processes of vowel epenthesis or deletion, vowel coloring by rounded and back consonants, and prevocalic stop voicing.

A sample form produced by the FST for the word *saabisbisdiithl* ‘they tore off (pl. common noun)’ is in example (7). This form involves a preverb *saa* being affixed directly to a transitive verb *bisbis*, a reduplicated plural form of the verb which was listed directly in the H&R wordlist (the symbol ^ marks morpheme boundaries).[3] After the verb, we find two inflectional suffixes and one clitic. Ultimately, rewrite rules are used to delete the transitive suffix and segmentation boundaries (8).

(7)
```
saa^bisbis^i^diit^hl
saa+PVB-bisbis+VT-TR-3PL=CN
```

(8)
```
saabisbisdiithl
```

4.2 Analyzer iterations

We built and evaluated four iterations of the Gitksan morphological analyzer based upon the foundation presented in Section 4.1: the v1. **Lexical FST**, v2. **Complete FST**, v3. **Dialectal FST** and v4. **FST+Neural**. Each iteration cumulatively expands the previous one by incorporating additional vocabulary items, rules or modeling components.

The first analyzer (v1: **Lexical FST**) included only the open-class categories of verbs, nouns, modifiers, and adverbs which made up the bulk of the H&R wordlist. The main focus of the morphotactic description was transitive inflection, person/number-agreement, and cliticization for these categories. Some semi-productive argument structural morphemes (e.g. the passive *-xw* or antipassive *-asxw*) were also included.

The second analyzer (v2: **Complete FST**) incorporated functional and closed-class morphemes such as subordinators, pronouns, prepositions, quotatives, demonstratives, and aspectual particles, including additional types of clitics.

The third analyzer (v3: **Dialectal FST**) further incorporated predictable stem-internal variation, such as the vowel shift and dorsal stop lenition/-fortition seen across dialects. In order to apply the vowel shift in a targeted way, all items in the lexicon were marked for stress using the notation `$`. Parses prior to rule application now appear as in (9) (compare to (7)).

(9)
```
s$aa^bisb$is^i^diit^hl
```

Finally, we seek to expand the coverage of the analyzer through machine learning, namely neural architectures (v4: **FST+Neural**). Our FST architecture allows for the automatic extraction of surface-analysis pairs; this enables us to create

[3] The FST has no component to productively handle reduplication but this would be possible to implement given a closed lexicon Hulden (2009a, Ch. 4).

```
LEXICON RootN
maa'y      N ;
smax       N ;
LEXICON RootVI
yee        VI ;
t'aa       VI ;
LEXICON RootPrenoun
lax_       Prenoun ;
```

(a) Lexicon

```
LEXICON N
+N:          NInfl ;
LEXICON NInfl
-ATTR:^m     # ;
-SX:^it      Word ;
             Agr_II ;
             Word ;
LEXICON Prenoun
+PNN:        # ;
+PNN:        RootN ;
```

(b) Morphotactic description

Deletion before -3PL:
ˆi → 0 / _ ˆdiit

Vowel insertion:
0 → i / C ˆ _ Sonorant #

Prevocalic voicing:
p,t,ts,k,$\underline{k}$ → b,d,j,g,$\underline{g}$ / _ V

(c) Rewrite rules

Figure 3: Three main components of the FST (simplified)

a training set for the neural models. We experiment with two alternative neural architectures - the Hard-Attentional model over edit actions (**HA**) described by Makarov and Clematide (2018), and the transformer model (Vaswani et al., 2017), as implemented in Fairseq (**Fairseq**) (Ott et al., 2019). Unlike the FST, the neural models can extend morphological patterns beyond a defined series of stems, analyzing forms that the FST cannot recognize.

For both models, we extract 10,000 random analysis pairs, with replacement; early stopping for both models uses a 10% validation set extracted from the training, with no overlap between training and validation sets (although stem overlap is allowed). The best checkpoint is chosen based on validation accuracy. The HA model uses a Chinese Restaurant Process alignment model, and is trained for 60 epochs, with 10 epochs patience; the encoder and decoder both have hidden dimension 200, and are trained with 50% dropout on recurrent connections. The Transformer model is a 3-layer, 4-head transformer trained for 50 epochs. The encoders and decoders each have an embedding size of 512, and feed-forward size of 1024, with 50% dropout and 30% attentional dropout. We optimize using Adam (0.9, 0.98), and cross-entropy with 20% label-smoothing as our objective.

Any wordform which received no analysis from the FST was provided a set of five possible analyses each from the HA and Fairseq models.

5 Evaluation

5.1 FST Coverage

The analyzers were run on two 2000-token datasets drawn from the multidialectal corpus: an Eastern Gitksan dataset (1 speaker), and a Western Gitksan dataset (2 speakers and dialects). Token and type coverage for the three FSTs is provided in Table 1, representing the percentage of wordforms for which each analyzer was able to provide one or more possible parses.

		Types	Tokens
East	Lexical	63.12%	54.17%
	Complete	71.10%	81.48%
	Dialectal	71.10%	81.48%
West	Lexical	45.49%	38.09%
	Complete	53.20%	70.12%
	Dialectal	62.35%	75.98%

Table 1: Analyzer coverage on 2000-token datasets

The effect of adding function-word coverage to the second 'Complete' analyzer was broadly similar across dialects, increasing type coverage by about 8% and token coverage by 27-32%, demonstrating the relative importance of function words to lexical coverage.

The first two analyzers performed substantially better on the Eastern dataset which more closely matched the dialect of the wordlist/lexicon. The third 'Dialectal' analyzer incorporated four types of predictable stem-internal allomorphy to generate Western-style variants. These transformations had no effect on coverage for the Eastern dataset, but increased type and token coverage for the Western dataset by 9% and 6% respectively.

5.2 FST precision

While our analyzer manipulates vocabulary items at the level of the stem seen in the lexicon, the corpus used for evaluation is annotated to the level of the root and was not always comparable (e.g. *ihlee'etxw* 'red' vs *ihlee'e-xw* 'blood-VAL'). Accuracy evaluation therefore had to be done manually by comparing the annotated analysis in the corpus to the parse options produced by the FST (10).

(10) *japhl*

a. `make[-3.II]=CN` (Corpus)

b. `j$ap+N=CN`
`j$ap+N-3=CN`
`j$ap+VT-3=CN` (FSTv3)

We evaluated the accuracy of the Dialectal FST on two smaller datasets: 150 tokens Eastern, and 250 tokens Western. These datasets included 85 and 180 unique wordform/annotation pairs respectively. The same wordform might have multiple attested analyses, depending on its usage. The performance of the Dialectal analyzer on each dataset is summarized in Table 2. Precision is calculated as the percentage of word/annotation pairs for which the analyzer produced a parse matching the context-sensitive annotation in the corpus.[4] Other analyses produced by the FST were ignored. For example in (10), the token would be evaluated as correct given the final parse, which uses the appropriate stem (*jap* ‘make’) and matching morphology; the other parses using a different stem (*jap* ‘box trap’) and/or different morphology could not qualify the token as correctly parsed. Only parsable wordforms were considered (i.e. English words and names are excluded).

	East	West
Coverage	71.76% (61/85)	68.89% (124/180)
Correct parse	71.76% (61)	64.44% (116)
Incorrect parse	0.00% (0)	2.78% (5)
Name, English	2.5% (2)	3.33% (6)
No parse	27.5% (22)	29.44% (53)
Precision	100.00% (61/61)	95.87% (116/121)

Table 2: Accuracy evaluation for dialectal analyzer (v3) on small datasets

The Western dataset was larger, and consisted of two distinct dialects, in contrast to the smaller and more homogeneous Eastern dataset. Regardless, analyzer coverage between the two datasets was comparable (68-72%) and precision was very high (95-100%). When this analyzer was able to provide a possible parse, one was almost always correct.

To further understand the analyzer’s limitations, we categorized the reasons for erroneous and missing analyses, listed in Table 3. In addition to the small datasets, for which all words were checked, we also evaluated the 100 most-frequent word/analysis pairs in the larger datasets.

The majority of erroneous and absent analyses were due to the use of new lemmas not in the lexicon, or novel variants not captured by productive stem-alternation rules. Novel lemmas made up about 18% each of the small datasets, and 4-8% of the top-100 most frequent types. Some functional items had specific dialectal realizations; for example, all three speakers used a different locative preposition (*g̲oo-*, *g̲o’o-*, *g̲a’a-*), only one of which was recognized.

There were also a few errors attributable to the morphotactic rules encoded in the parser. For example, there were several instances in the dataset of supposed ‘preverb’ modifiers combining with nouns (e.g. *t’ip-no’o=si*, sharply.down-hole=PROX, ‘this steep-sided hole’), which the parser could not recognize. This category combination flags the need for further documentation of certain ‘preverbs’. As a second example, numbers attested without agreement were not recognized because the analyzer expected that they would always agree. This could be fixed by updating the morphotactic description for numbers (e.g. to more closely match intransitive verbs).

5.3 FST + Neural performance

The addition of the neural component significantly increased the analyzer’s coverage (mean HA: +21%, Fairseq: +17%), but at the expense of precision (mean -15% for both). The results of the manual accuracy evaluation are presented in Figure 4. There remained several forms for which the neural analyzers produced no analyses.

Both analyzers performed better on the 100-most-frequent types datasets, where they tended to accurately identify dialectal variants of common words (e.g. *t’ihlxw* from *tk’ihlxw* ‘child’, *diye* from *diya* ‘3=QUOT (third person quotative)’). In the small datasets of running text, these models were occasionally able to correctly identify unknown noun and verb stems that had minimal inflection. However, they struggled with identifying categories, and often failed to identify correct inflection. These difficulties stem from category-flexibility and homophony in Gitksan. Nouns and

[4] Note that precision is computed only on word forms which received at least one analysis from the FST.

	East		West	
	150 tokens (22)	Top-100 (17)	250 tokens (58)	Top-100 (23)
New lemma	15	2	30	2
New function word	1	2	4	6
Lexical variant	3	8	6	5
Functional variant	2	3	9	9
Morphotactic error	1	2	9	1

Table 3: Categorization of erroneous and absent analyses for dialectal analyzer (FSTv3)

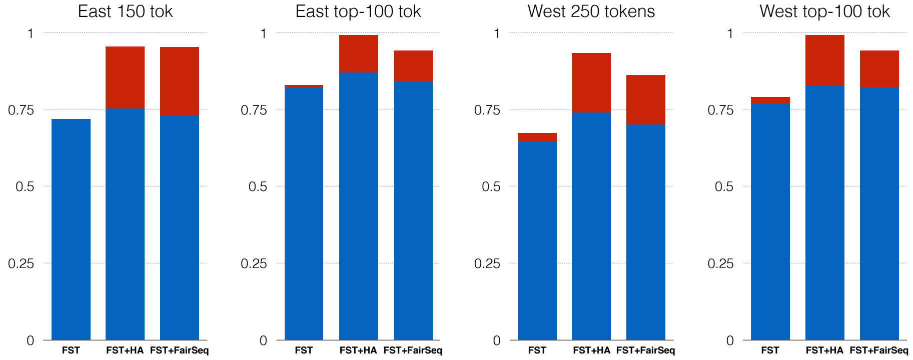

Figure 4: Proportion of forms which receive the correct analysis from each of our models (indicated in blue) and the number of forms which receive only incorrect analyses from our models (indicated in red). The remaining forms received no analyses.

verbs use the exact same inflection and clitics, making the category itself difficult to infer. Short inflectional sequences have a large number of homophonous parses, and even more differ only by a character or two.

Qualitatively, the HA model tended to produce more plausible predictions, often producing the correct stem or else a mostly-plausible analysis that could map to the same surface form, but with incorrect categories or inflection. In contrast, the Fairseq model often introduced stem changes or inflectional sequences which could not ultimately map to the surface form. Example (11) provides a sample set of incorrect predictions (surface-plausible analyses are starred).

(11) *ksimaasdiit* `ksi+PVB-m$aas+VT-TR-3PL`

a. HA model
`xsim$aas+N-3PL` (*)
`xsim$aas+N-T-3PL` (*)
`xsim$aas+NUM-3PL` (*?)
`xsim$aast+N-T-3PL`

b. Fairseq model
`xsim$aast+N-3PL`
`xsim$aast+N=RESEM`
`xsim$aast+N-SX=PN`
`xsim$as+N-3PL`

Further work can be done to improve the performance of the neural addition, such as training the model on attested tokens instead of, or in addition to, tokens randomly generated from the FST analyzer.

6 Discussion and Conclusions

The grammatically-informed FST is able to handle many of Gitksan's morphological complexities with a high degree of precision, including homophony, contextual deletion, and position-flexible clitics. The FST analyzer's patchy coverage can be attributed to its small lexicon. Unknown lexical items and variants comprised roughly 18% of each small dataset. Notably, errors and unidentified forms in the FST analyzer signal the current limits of morphotactic descriptions and lexical documentation. The analyzer can therefore serve as a useful part of a documentary linguistic workflow to quickly and systematically identify novel lexical items and grammatical rules from texts, facilitating the expansion of lexical resources. It can also be used as a pedagogical tool to identify word stems in running text, or to generate morphological exercises for language learners.

The neural system, with its expanded coverage, can serve as part of a feedback system with a human in the loop, informing future iterations of the annotation process. While its precision is lower than the FST, it can still inform annotators on words that the FST does not analyze. Newly-annotated data can then be used to enlarge the FST coverage.

Acknowledgments

'Wii t'isim ha'miyaa nuu'm aloohl the fluent speakers who continue to share their knowledge with me (Barbara Sennott, Vincent Gogag, Hector Hill, Jeanne Harris), as well as the UBC Gitksan Research Lab. This research was supported by funding from the National Endowment for the Humanities (Documenting Endangered Languages Fellowship) and the Social Sciences and Humanities Research Council of Canada (Grant 430-2020-00793). Any views/findings/conclusions expressed in this publication do not necessarily reflect those of the NEH, NSF or SSHRC.

References

Kenneth R Beesley and Lauri Karttunen. 2003. Finite-state morphology: Xerox tools and techniques. *CSLI, Stanford*.

Colin Brown, Clarissa Forbes, and Michael David Schwan. 2020. Clause-type, transitivity, and the transitive vowel in Tsimshianic. In *Papers of the International Conference on Salish and Neighbouring Languages 55*. UBCWPL.

Emily Chen and Lane Schwartz. 2018. A morphological analyzer for st. lawrence island/central siberian yupik. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Henry Davis. 2018. Only connect! a unified analysis of the Tsimshianic connective system. *International Journal of American Linguistics*, 84(4):471–511.

Britt Dunlop, Suzanne Gessner, Tracey Herbert, and Aliana Parker. 2018. Report on the status of BC First Nations languages. Report of the First People's Cultural Council. Retrieved March 24, 2019.

Clarissa Forbes. 2018. *Persistent ergativity: Agreement and splits in Tsimshianic*. Ph.D. thesis, University of Toronto.

Clarissa Forbes, Henry Davis, Michael Schwan, and the UBC Gitksan Research Laboratory. 2017. Three Gitksan texts. In *Papers for the 52nd International Conference on Salish and Neighbouring Languages*, pages 47–89. UBC Working Papers in Linguistics.

Atticus G Harrigan, Katherine Schmirler, Antti Arppe, Lene Antonsen, Trond Trosterud, and Arok Wolvengrey. 2017. Learning from the computational modelling of plains cree verbs. *Morphology*, 27(4):565–598.

Lonnie Hindle and Bruce Rigsby. 1973. A short practical dictionary of the Gitksan language. *Northwest Anthropological Research Notes*, 7(1).

Mans Hulden. 2009a. *Finite-state machine construction methods and algorithms for phonology and morphology*. Ph.D. thesis, The University of Arizona.

Mans Hulden. 2009b. Foma: a finite-state compiler and library. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics: Demonstrations Session*, pages 29–32. Association for Computational Linguistics.

Benjamin Hunt, Emily Chen, Sylvia L.R. Schreiner, and Lane Schwartz. 2019. Community lexical access for an endangered polysynthetic language: An electronic dictionary for St. Lawrence Island Yupik. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 122–126, Minneapolis, Minnesota. Association for Computational Linguistics.

Katharine Hunt. 1993. *Clause Structure, Agreement and Case in Gitksan*. Ph.D. thesis, University of British Columbia.

Lingpeng Kong, Chris Dyer, and Noah A Smith. 2015. Segmental recurrent neural networks. *arXiv preprint arXiv:1511.06018*.

Patrick Littell. 2018. Finite-state morphology for kwak'wala: A phonological approach. In *Proceedings of the Workshop on Computational Modeling of Polysynthetic Languages*, pages 21–30, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Patrick Littell, Aidan Pine, and Henry Davis. 2017. Waldayu and waldayu mobile: Modern digital dictionary interfaces for endangered languages. In *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 141–150.

Peter Makarov and Simon Clematide. 2018. Neural transition-based string transduction for limited-resource setting in morphology. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 83–93, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Jeffrey Micher. 2017. Improving coverage of an Inuktitut morphological analyzer using a segmental recurrent neural network. In *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 101–106, Honolulu. Association for Computational Linguistics.

Sarah Moeller, Ghazaleh Kazeminejad, Andrew Cowell, and Mans Hulden. 2018. A neural morphological analyzer for arapaho verbs learned from a finite state transducer. In *Proceedings of the Workshop on Computational Modeling of Polysynthetic Languages*, pages 12–20.

Mother Tongues Dictionaries. 2020. Gitksan. Edited by the UBC Gitksan Research Lab. Accessed June 4, 2020. (`https://mothertongues.org/gitksan`).

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Bruce Rigsby. 1986. Gitxsan grammar. Master's thesis, University of Queensland, Australia.

Lane Schwartz, Emily Chen, Benjamin Hunt, and Sylvia LR Schreiner. 2019. Bootstrapping a neural morphological analyzer for st. lawrence island yupik from a finite-state transducer. In *Proceedings of the Workshop on Computational Methods for Endangered Languages*, volume 1.

Conor Snoek, Dorothy Thunder, Kaidi Loo, Antti Arppe, Jordan Lachler, Sjur Moshagen, and Trond Trosterud. 2014. Modeling the noun morphology of plains cree. In *Proceedings of the 2014 Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 34–42.

Tonya Stebbins. 2003. On the status of intermediate form-classes: Words, clitics, and affixes in Coast Tsimshian (Sm'algyax). *Linguistic Typology*, 7(3):383–415.

Lonny Strunk. 2020. *A Finite-State Morphological Analyzer for Central Alaskan Yup'Ik*. Ph.D. thesis, University of Washington.

Marie-Lucie Tarpent. 1987. *A Grammar of the Nisgha Language*. Ph.D. thesis, University of Victoria.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762*.