Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 6597–6606
Marseille, 20-25 June 2022

An Inflectional Database for Gitksan

Bruce Oliver[ϒ], Clarissa Forbes[Φ], Changbing Yang[ϒ], Farhan Samir[ϒ], Edith Coates[ϒ], Garrett Nicolai[ϒ], Miikka Silfverberg[ϒ]
[ϒ]University of British Columbia, [Φ]Independent
bruce.oliver@ubc.ca, forbesc@alumni.ubc.ca, cyang33@student.ubc.ca, fsamir@mail.ubc.ca
icoates1@mail.ubc.ca, garrett.nicolai@mail.ubc.ca, msilfver@mail.ubc.ca

Abstract

This paper presents a new inflectional resource for Gitksan, a low-resource Indigenous language of Canada. We use Gitksan data in interlinear glossed format, stemming from language documentation efforts, to build a database of partial inflection tables. We then enrich this morphological resource by filling in blank slots in the partial inflection tables using neural transformer reinflection models. We extend the training data for our transformer reinflection models using two data augmentation techniques: data hallucination and back-translation. Experimental results demonstrate substantial improvements from data augmentation, with data hallucination delivering particularly impressive gains. We also release reinflection models for Gitksan.

Keywords: Gitksan, computational morphology, low-resource, reinflection

1. Introduction

In recent years, research on computational morphology has enjoyed access to an ever-increasing variety of annotated data through the Universal Morphology (UniMorph) database (McCarthy et al., 2020)—a collection of inflection tables which now covers hundreds of languages (see Figure 1 for an example). Most of the data in UniMorph stems from Wiktionary and is automatically converted into UniMorph format. Wiktionaries are, however, available only for a limited number of languages. In this paper, we explore an alternative route for generation of high-quality paradigm-level inflectional data. Our starting-point is interlinear glossed text (IGT) for the Tsimshianic language Gitksan, generated in language documentation efforts. We convert this text into a paradigm-level inflectional resource using a combination of rule-based methods and neural morphological reinflection (Cotterell et al., 2016).

IGT is a semi-structured representation format which is commonly employed in language documentation. It consists of transcribed utterances paired with morphological segmentation information, morphosyntactic annotations, and translations into a meta language (English in our case). See Figure 2 for a Gitksan example. IGT is a useful starting point for construction of morphological databases because of its extensive morphological annotation.

Conversion of IGT into inflection tables entails many design decisions. For example, due to the underdocumented status of Gitksan, it is not clear which morphological processes constitute inflection and which are more properly classified as derivation. Section 5 presents our criteria for distinguishing between Gitksan inflection and derivation along with the conversion pipeline from IGT to inflection tables.

Our paper is accompanied by the first database of inflection tables for Gitksan. Since not all inflected forms of every lexeme are attested in our underlying IGT resource, we additionally release neural reinflection models for filling in missing forms in the tables.[1]. The final inflectional database has a multitude of applications: it can be used for construction of language learning resources, experiments in computational morphology and training of automatic systems for glossing, which can speed up the documentation effort for Gitksan.

outran	V;PST
outrunning	V;V.PTCP;PRS
outruns	V;3;SG;PRS
outrun	V;NFIN
outrun	V;V.PTCP;PST

Figure 1: An English morphological paradigm from the UniMorph database which lists all inflected forms of the lexeme /outrun/ together with their morphosyntactic descriptions.

In order to fill in empty slots in our partial inflection tables, we treat incomplete tables as an instance of the paradigm cell-filling problem (PCFP) (Ackerman et al., 2009), training transformer models for inflection generation, and applying the models to predict missing forms. To combat model overfitting caused by data scarcity, we experiment with data augmentation using hallucination (Anastasopoulos and Neubig, 2019) and back-translation for inflection (Sennrich et al., 2016; Liu and Hulden, 2021a). Data hallucination delivers substantial gains in inflection accuracy and our final models achieve 73.95% full form accuracy in the challenging split-by-lemma setting (Goldman et al., 2021b), where none of the lexemes represented in the training data occur in the test data. This represents an absolute improvement of 16%-points in inflection accuracy over a baseline transformer model

[1] https://github.com/mpsilfve/gitksan-data

which does not make use of data augmentation.

2. The Gitksan Language

The Gitxsan are one of the indigenous peoples of the northern interior region of British Columbia, Canada. Their traditional territories consist of upwards of 50,000 square kilometers of land in the upriver Skeena River watershed area. Their traditional language, called Gitksan in the linguistic literature, 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). While Gitksan is still in use in formal settings such as feasts, it is quickly reaching a point of turnover, with cultural duties falling to generations that are not fluent. Community revitalization efforts are underway but are primarily undertaken by individuals on an ad-hoc basis. Initiatives include regular in-school language programming, a few adult language courses, a successful language immersion camp, and several Master-Apprentice pairs. Linguistic documentation on Gitksan and the Tsimshianic languages has been going on intermittently since the 1970s, including the drafting of a never-published grammar (Rigsby, 1986) and waves of formal phonological, syntactic, and semantic work over the past thirty years. There are several community-developed word lists and grammar workbooks, but no comprehensive dictionary, grammar, or pedagogical materials.

Gitksan, along with all its Tsimshianic relatives, has fairly strict VSO word order and "analytic to synthetic" morphology (Rigsby, 1986; Rigsby, 1989). That is, unlike many Canadian indigenous languages, it is not polysynthetic, making paradigm-generation tasks feasible to undertake without reliance on sub-word models. Like many languages of the Americas, Gitksan exhibits head-marking (agreement), rather than dependent-marking (case). It has a rich assortment of derivational morphemes and substantial capacity for compounding; consequently, its degree of word-complexity has been described as similar to German (Tarpent, 1987).

3. Related Work

Paradigm Cell-Filling Problem The paradigm cell-filling problem (Ackerman et al., 2009) (PCFP) is an inflection task, where unseen inflected forms of a lexeme are generated with known forms of the lexeme as input. Many effective systems for the PCFP task have been developed (Silfverberg and Hulden, 2018; Kann et al., 2017; Cotterell et al., 2017).

Morphology resources The idea of using IGT as a basis for morphological databases is not novel. Moeller et al. (2020) present a general approach for extracting paradigms from interlinear glossed text and completing missing forms in the resulting partial paradigms using deep sequence-to-sequence models. Apart from minor details like choice of underlying machine learning architecture, our approach is similar. However, whereas Moeller et al. (2020) distinguish between inflectional and derivational processes in only the most obvious cases, we carefully filter out all derivational material from our inflection tables leading to a more restricted selection of forms which is better suited for construction of language learning applications.

UniMorph (Kirov et al., 2018) is the most widely used database for inflectional morphology. Universal Dependencies treebanks (Nivre et al., 2016) provide another resource for morphologically analyzed word forms. However, unlike UniMorph, Universal Dependencies data will contain only a small subset of forms for a given lexeme. It also does not stringently distinguish between inflectional and derivational processes.

Our resource is not the first resource of morphological paradigms for Indigenous languages of the Americas: Cruz et al. (2020) present a dataset of inflectional verb paradigms for Chatino.

Computational Morphology for Indigenous Canadian Languages Many finite-state morphological analyzers have been created for Canadian Indigenous languages in recent years: Yupik (Strunk, 2020; Chen and Schwartz, 2018), Plains Cree (Harrigan et al., 2017; Snoek et al., 2014), Kwak'wala (Littell, 2018), Gitksan (Forbes et al., 2021), and others. A number of authors have also investigated hybrid finite-state and data-driven neural models of morphology for Indigenous Canadian languages (Schwartz et al., 2019; Moeller et al., 2018; Kong et al., 2015).

Computational Work for Gitksan 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 frequently occurs in Gitksan.

4. Gitksan IGT Data

The starting-point for this project is a corpus of interlinear-glossed narratives, extended from an initial collection of three stories presented by Forbes et al. (2017). The dataset consists of a total of 12,613 tokens (2,472 unique types). The oral narratives come from three speakers from three distinct dialect areas, in roughly equal amounts: Ansbayaxw (Eastern); Gijigyukwhla'a and Git-anyaaw (Western). These are transcribed in the accepted community orthography (Hindle and Rigsby, 1973), in a manner that orthographically represents speaker and dialectal variation, and include free translations from the speakers. The texts are annotated in an interlinear-gloss format based on the Leipzig Glossing Rules (Max Planck Institute and University of Leipzig, 2008), with language-specific additions, as illustrated in 2. The first line in the figure is an orthographic representation, the second line is a

Orthography	Ii	sagootxwt	dimt	wila	liluxwshl	hun.
Segmentation	ii	sa-goot-xw-t	dim=t	wila	liluxws[-t]=hl	hun
Gloss	CCNJ	CAUS1-heart-VAL-3.II	PROSP=3.I	MANR	steal[-3.II]=CN	salmon
Translation	And he planned to steal a fish.					

Figure 2: An example interlinear-glossed sentence demonstrating the corpus annotation format and typical Gitksan morphological structure. The second line includes segmentations with morphemes normalized to a canonical orthographic form. The third line has an abbreviated gloss for each segmented morpheme.

segmentation of the orthographic form, with both inflectional and derivational morphemes segmented apart from a recognizable etymological root. The third line is a gloss for each of the segmented morphemes, with abbreviations for functional morphemes presented in uppercase. The data additionally contains English translations.

An important note about IGT is that it is very time-consuming to produce, particularly larger collections like this one, and that this one is still being edited. The inflectional paradigms that we present are based on an initial draft of the collection after annotations were completed. However, several major revisions to the collection have since been made: some are orthographic decisions, some are changes to the annotation format and segmentation style, and some are analytical corrections and typo fixes. We intend to produce a second version of our inflection tables once the IGT collection is finalized.

5. Conversion into Morphological Paradigms

This section presents our conversion pipeline from IGT to inflectional paradigms. Our aim is to identify inflected noun and verb forms in the IGT and group them into tables according to lexeme.

5.1. Defining the Structure of Inflection Tables

We will first present the inventory of inflectional types in our tables. This is a subset of all forms occurring in the IGT resource. We include only inflectional morphology in the tables, whereas the IGT resource also contains a rich variety of derivational morphology and a number of clitic affixes. For Gitksan, there are no established criteria for grouping morphological processes into inflectional vs. derivational processes. We, therefore, decided to apply the following four criteria to identify inflectional affixes:

- An inflectional affix should never change the lemma's part of speech, such as by transforming a noun into a verb or vice versa.[2]
- An inflectional affix should apply widely to most lexemes of the same part of speech category.
- An inflectional affix should convey a well-defined syntactic/semantic function.
- An inflectional affix should occur frequently in the IGT resource.

Some examples of affixes excluded by these criteria are given in Table 1. When possible, we treat derived forms as their own lexemes, adding a separate table for inflectional variants of the derived form. For example, the derived form *sagootxw* 'idea, decision, plan' receives its own inflection table, separate from the inflection table for its root lemma *goot* 'heart'. This substantially increases the number of inflection tables in our dataset. The remaining affix combinations determined the form inventory in our inflection tables. We were left with 33 forms which are described in Appendix A.

Our inflection tables are similar across categories because nouns and verbs in Gitksan inflect similarly, for example using the same agreement for absolutive and possessors.[3]

5.2. Conversion of inflected forms

Next, we populate inflection tables with inflected forms. This process is illustrated in Figure 3. Each word form in an utterance in the IGT resource like *liluxwshl* is accompanied by two distinct annotations: a morphological segmentation *liluxws[-t]=hl* which lists its component morphemes and a gloss *steal[-3.II]=CN* which contains an English translation of the word root, *steal* in this case, and the morphosyntactic tags of each bound morpheme in the word. The morphological segmentation is not a pure segmentation in the sense that the component morphemes are normalized into canonical forms.

Form filtering As a preprocessing step, we filter out all words which are not noun or verb forms. Because there are no part-of-speech tags in the IGT resource, we identify these based on their English glosses, when these are nouns and verbs.

Stem identification We start the actual conversion process by identifying the stem of the word. For many word forms, this will be the root morpheme

[2]Note that the distinction between nouns and verbs is not clear-cut in Gitksan but prototypical examples of both do exist, which helped us apply this criterion.

[3]However, even though nouns, transitive verbs and intransitive verbs share most of their inflection, there are a number of forms which do not occur in all categories so there are minor differences between tables which are explained in Appendix B.

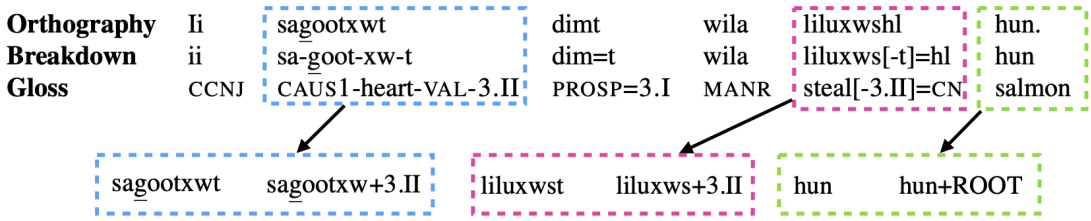


Figure 3: Illustration of how we extract infected forms and their morphosyntactic descriptions from interlinear glossed words. Key-points of the conversion pipeline include: (1) identification of stems like *sagootxw* which may contain derivational affixes such as the causative *sa-* and valence-shifting marker *-xw*, (2) removal of clitics like *=hl* and (3) recovery of hidden morphemes like *[-t]*. We also add *+ROOT* tags which signify the base form of a lexeme. Three of the words *Ii*, *dimt* and *wila* belong to closed classes and are excluded from our inflection tables.

Morpheme	Gloss	Rationale for exclusion	Example(s)
si-	Causative	Changes part of speech	*maa'y* 'berries (N)', *simaa'y* 'harvest berries (VI)'
-asxw	Antipassive	Not fully productive	*'wa* 'find (VT)', none or *'weesxw* 'be bountiful (VI)'
		Changes part of speech	*jakw* 'kill (VT)', *jagwasxw* 'animal (N)'
-xw	Valence marker	Function poorly defined	passive/intransitive: *kw'ootxw* 'be lost (VI)'
			transitivizer/applicative: *naksxw* 'marry (VT)'
			optional possessive: *sim'oogitxu'm* 'our chiefs'
na-	Reciprocal	Too few examples	*nadisityeexw(t)* 'trade with each other'

Table 1: Some examples of morphemes which were excluded from the inflection tables on the basis of being (1) not sufficiently inflectional or (2) too infrequent/poorly understood to attempt to generate.

which is associated with the English translation. However, for words which contain derivational material like *sagootxwt* with segmentation *sa-goot-xw-t* and gloss *CAUS1-heart-VAL-3.II*, the stem will consist of the root with any derivational material, in this case *sagootxw*, consisting of the causative *sa* joined with the root *goot* followed by the valency marker *xw*. We utilize the morphological segmentation for determining multimorphemic stems. For each unique word stem, we create one inflection table.

Due to speaker and dialectal variation, the IGT resource contains a fair amount of allomorphic variation. For example, we see three different variants for the root 'people': *git*, *gat* and *get*. We decided to include all of these in the same inflection table. In order to avoid grouping together unrelated forms, this process was restricted to known alternation patterns.[4]

Clitic filtration Many forms in the IGT resource contain clitics which we do not include in our inflection tables. As an example, the form *liluxwshl* contains a clitic *hl* at the end. These are denoted in the Gitksan IGT resource using equals signs as in the example segmentation *liluxws[-t]=hl*. This allows us to easily identify and remove clitics in most cases. In some cases, the gloss will indicate that appending the clitic in fact deleted another morpheme. As an example, the square brackets *[..]* in the gloss *liluxws[-t]=hl* indicate that the morpheme *-t* was deleted. When deleting the clitic, we simultaneously restore this material to the word form. The form *liluxwshl* with gloss *steal[-3.II]=CN* is therefore converted into *liluxwst* with gloss *steal-3.II* and then added into the inflection table for the stem *liluxws*.

[4]For example, the *a/e* alternation in *gat/get* between the Eastern and Western dialects is well known.

5.3. Paradigm-Level Resource

We will now describe the final resource of inflection tables. In total, there are 1055 inflection tables containing 2125 inflected forms. Only 5.3% of the slots in our tables contain a form and 51% of all tables contain only a single filled slot. The rest of the forms are not attested in the IGT resource. Of the slots which do contain forms, 13.4% contain more than one form. These are dialectal and spelling variants. For example, there are three distinct base forms for the lexeme /gat/ 'people': *git*, *gat* and *get*.

6. Experiments

Given the sparsity of our inflection tables, it is important to investigate methods to automatically generate forms in empty table slots. We now describe experiments on paradigm cell-filling. Here we train morphological reinflection models to predict missing slots in the inflection tables, thereby completing partial tables. We investigate the effect of different data augmentation techniques in an attempt to improve the accuracy of our inflectors.

6.1. Data for PCFP Experiments

Data Split We train and test models on the Gitksan morphological paradigms which were described in

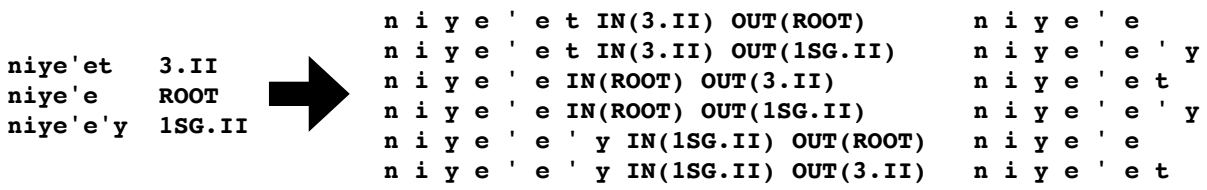


Figure 4: From a paradigm in the training data spanning three forms, we can generate six reinflection training examples. Forms are split into individual characters and we distinguish tags for input forms from tags for output forms.

Data	Train	Dev	Test
Inflected forms	858	302	124

Table 2: Data splits for the reinflection experiment

Section 5. As noted, these paradigms are very sparse, with most paradigms containing a single form and the overwhelming majority containing a handful of forms. For our experiments, we first discard all tables which contain a single form.[5]. We additionally retain only a single form if there are multiple possible forms in one table slot (multiple possible realizations of a particular slot might occur for example due to dialectal variation). In this case, we randomly sample one among the alternatives given in the table.

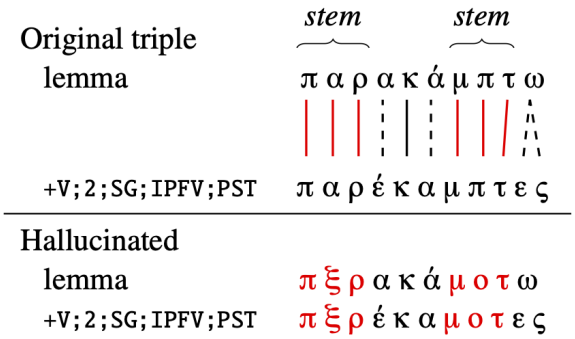
We set aside 17 tables for testing and use the remaining tables for model training and fine-tuning. The decision to evaluate purely on tables which were not observed during training reflects recent observations that evaluation on lexemes, which were observed during training, can lead to gross exaggeration of inflection performance, especially in low-resource settings (Liu and Hulden, 2021b; Goldman et al., 2021a). All data details are shown in Table 2.

Training data In order to train PCFP models, we generate reinflection examples from all tables in the training data. For any pair of forms f_1 and f_2 belonging to the same inflection table, we generate a reinflection training example $f_1 \mapsto f_2$. Similarly, we produce reinflection examples for the development data. This corresponds to the 1-source data format for reinflection presented by Liu and Hulden (2020). See Figure 4 for details.

6.2. Model Architecture

We train all models using the Fairseq (Ott et al., 2019) implementation of transformers (Vaswani et al., 2017). Both the encoder and decoder have 4 layers with 4 attention heads, an embedding size of 256 and hidden layer size of 1024. We train with the Adam optimizer starting of the learning rate at 0.001. We chose the batch size (400) and maximum updates (20, 000) based on preliminary experiments on the held-out development set. Our model setting resembles the work of Wu et al. (2021) who found that a relatively large batch size could benefit morphological inflection. Prediction is performed with the best checkpoint model, according to validation accuracy, and we use a beam of width 5 during inference.

[5]As explained below, neither train nor test examples can be extracted from tables containing a single form. We can predict missing forms but there is no way to evaluate accuracy without providing additional forms manually.



[Illustration from Anastasopoulos and Neubig (2019)]

Figure 5: Illustration of the data hallucination approach where the longest common subsequence in the input and output example are replaced with a random character string.

6.3. Data augmentation Techniques

Data Hallucination Data sparsity causes problems in low data conditions for encoder-decoder models. They have a tendency to predict common character sequences which occur in the training data while ignoring the input example, a problem which has been dubbed label bias (Wiseman and Rush, 2016). In order to address label bias, we employ data hallucination, augmenting our limited training data with synthetic examples. We employ the approach proposed by Anastasopoulos and Neubig (2019), where noise is introduced into existing training examples by replacing the longest common subsequence of input and output forms with random character strings. The approach is illustrated in Figure 5. We experiment with different sizes for the augmented dataset, adding 150, 300, 650, 1,350 or 1,950 synthetic examples to the original dataset. We then tune the number of synthetic examples, maximizing inflection accuracy on our held-out development data. Based on the results, we use 1,350 synthetic examples for our final experiments.

Back-translation Another prominent method for data augmentation is back-translation: a technique for leveraging monolingual data in low-resource machine translation (Sennrich et al., 2016). It can also be applied in morphological inflection: the labeled data is leveraged to train a morphological inflection system to predict missing entries in paradigms. These predicted forms are then added to the original gold training data to train models for morphological inflection (Liu and Hulden, 2021a).

Concretely, we start by applying the baseline reinflection system to fill all empty slots in training tables. We then randomly sample a number of synthetic reinflection examples $f_1 \mapsto f_2$, where f_1 is a predicted form and f_2 is a gold standard form in the same table. The size of the sample examples starts from 150, 300, 650, 1,350, and up to 1,950. We experimented with the same augmented size as the data hallucination setting. Based on preliminary results on the development data, we combine $1,350$ synthetic examples with our original gold standard training data and train a reinflection system on this augmented dataset.

Tagged data augmentation Previous studies show that using a tag to distinguish between back-translated source sentences and gold standard training examples can improve translation performance (Caswell et al., 2019). Liu and Hulden (2021a) investigate this approach for inflectional generation task using back-translation, and find small benefits for some languages. Thus we also investigate this strategy both for hallucinated and back-translated synthetic training examples. A special tag <AUG> is appended at the end of the input of each synthetic training example. For example, the synthetic input:

```
x y z t IN(3.II) OUT(ROOT)
```

becomes:

```
x y z t IN(3.II) OUT(ROOT) <AUG>
```

6.4. Inference and Evaluation on Test Data

After training reinflection systems on the training data, we evaluate the systems on our test data consisting of 17 held-out inflection tables. For each table with n attested forms in the test data, we successively treat each attested form as a hidden output form and use the remaining $n-1$ forms in the table to predict the hidden form. This gives us n test cases for a table containing n filled slots.[6] We evaluate systems with respect to full-form accuracy.

When predicting a hidden test form in a table having n attested forms, we use $n-1$ forms as input. This gives us $n-1$ output candidates. We distill these down to a single model output based on one of two strategies:

1. **Random BL** We pick one of the forms at random.
2. **Majority** We apply majority voting.

Since the **Random BL** strategy includes a random component, we run inference three times with different random seeds and report average performance.

Augmentation strategy	Random BL	Majority
Baseline (none)	57.70	58.82
Data Hallucination	**72.55**	**73.95**
Data Hallucination+tags	68.06	71.43
Back-translation	54.62	55.46
Back-translation+tags	52.66	54.62

Table 3: Accuracy on test set.

7. Results

Quantitative evaluation The full form accuracy for all systems is presented in Table 3. Only one of the data augmentation strategies, namely data hallucination, results in an improvement over the baseline system, which does not employ data augmentation. Training with back-translated synthetic examples instead reduces performance by roughly 3%-points. We find that tagged data augmentation consistently hurts performance, with both hallucination and back-translation showing a small drop.

Choice of voting strategy is influential for model performance. Majority voting offers consistent benefits over the random strategy in all settings. Moreover, the error reduction provided by majority voting is greater for more powerful base-inflectors: For the baseline system, which does not employ data augmentation, majority voting offers an absolute improvement in accuracy of 1.1%-points, which corresponds to an error reduction of 2.6%. In contrast, for the best-performing base inflector which employs data hallucination, we get an absolute improvement in accuracy of 2.5%-points, corresponding to a 5.1% error reduction.

Error analysis In order to provide a more fine-grained analysis of the impact of data augmentation, we examine the differences between predictions generated by the baseline system and the system trained using our best-performing data augmentation setting, namely data hallucination with majority voting. We observe that most of the error reduction stems from correcting individual character omissions. For example, the baseline system predicts *hlibu* instead of the gold standard form *hlibuu* (thereby dropping one *u* character), but data hallucination corrects this omission.

We further examine the effect of tagged data augmentation on our best model: data hallucination with majority voting. It is hard to draw firm conclusions here, because the difference in accuracy for tagged and untagged hallucination is very small (71.43% vs. 73.95%). However, we do observe a tendency for the

[6] Note that this means that our test data is biased toward common inflected forms in the Gitksan IGT resource. This is likely to inflate the reported inflection accuracy to some extent.

tagged models to insert spurious characters in the output, for example predicting *t'aadin* (with an inserted *d*) instead of the gold standard form *t'aain*.

8. Discussion

Experiments Data augmentation seems to be crucial for inflection performance. Data hallucination improves performance from 58.82% to 73.95%. Nevertheless, performance is still not strong enough to allow for predicting missing slots in out tables without significant manual post-correction—a quarter of all predicted forms will have to be manually corrected. However, edit distance between the model output and gold standard form is small (1-2) in most cases, meaning that manual post-correction should usually be fast. In contrast to data hallucination, back-translation harms performance. This result is consistent with earlier work, which found a variable effect from back-translation (Liu and Hulden, 2021a). Similarly, tagged augmentation always under-performs untagged augmentation.

Dataset We confronted several challenges during the development of this dataset, especially when designing data conversion pipelines and inflectional table structures during data preprocessing. Consultation with expert linguists was crucial to enable necessary preprocessing steps such as distinguishing inflectional and derivational affixes, noting instances of dialect variation and other orthographic alternations, and identifying homophones and forms with identical glosses.
Although we have contributed and expanded a new data resource within this paper, there are still elements that need to be taken into consideration given the status of documentary resources as 'ongoing'. The IGT itself is undergoing revision, and dictionary resources which include part of speech information are under construction. Once both of these resources are further advanced, our inflection tables can be re-generated using the updated dataset, and with the structure of inflection tables customized to contain cells only relevant to each lemma's part of speech. These updates, particularly regarding the part of speech, are necessary for later application of the inflection tables for construction of language learning resources.

9. Conclusions

This paper presented a new resource of inflection tables for the Indigenous language Gitksan, and experimented with enriching the tables by predicting missing forms using transformer inflection models. For our experiments we applied two data augmentations techniques: data hallucination and back-translation. One of the techniques, namely data hallucination, improves prediction accuracy substantially. In future work, we aim to improve the resource by adding POS tags and revising the inflection tables according to forthcoming revisions to the Gitksan IGT data which forms the basis of the present project.

10. Acknowledgements

We want to thank Henry Davis, Lisa Matthewson and the Gitksan research lab at the Department of Linguistics at UBC for generous help with this project and access to Gitksan IGT data. We also want to thank for anonymous reviewers for valuable comments. 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.

11. Bibliographical References

Ackerman, F., Blevins, J. P., and Malouf, R. (2009). Parts and wholes: Implicative patterns in inflectional paradigms. *Analogy in grammar: Form and acquisition*, pages 54–82.

Anastasopoulos, A. and Neubig, G. (2019). Pushing the limits of low-resource morphological inflection. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 984–996, Hong Kong, China, November. Association for Computational Linguistics.

Caswell, I., Chelba, C., and Grangier, D. (2019). Tagged back-translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 53–63, Florence, Italy, August. Association for Computational Linguistics.

Chen, E. and Schwartz, L. (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)*.

Cotterell, R., Kirov, C., Sylak-Glassman, J., Yarowsky, D., Eisner, J., and Hulden, M. (2016). The sigmorphon 2016 shared task—morphological reinflection. In *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 10–22.

Cotterell, R., Sylak-Glassman, J., and Kirov, C. (2017). Neural graphical models over strings for principal parts morphological paradigm completion. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 759–765.

Cruz, H., Anastasopoulos, A., and Stump, G. (2020). A resource for studying chatino verbal morphology. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2827–2831, Marseille, France, May. European Language Resources Association.

Dunlop, B., Gessner, S., Herbert, T., and Parker, A. (2018). Report on the status of BC First Nations lan-

guages. Report of the First People's Cultural Council. Retrieved March 24, 2019.

Forbes, C., Davis, H., Schwan, M., 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.

Forbes, C., Nicolai, G., and Silfverberg, M. (2021). An FST morphological analyzer for the Gitksan language. In *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 188–197, Online, August. Association for Computational Linguistics.

Goldman, O., Guriel, D., and Tsarfaty, R. (2021a). (un) solving morphological inflection: Lemma overlap artificially inflates models' performance. *arXiv preprint arXiv:2108.05682*.

Goldman, O., Guriel, D., and Tsarfaty, R. (2021b). (un)solving morphological inflection: Lemma overlap artificially inflates models' performance. *ArXiv*, abs/2108.05682.

Harrigan, A. G., Schmirler, K., Arppe, A., Antonsen, L., Trosterud, T., and Wolvengrey, A. (2017). Learning from the computational modelling of plains cree verbs. *Morphology*, 27(4):565–598.

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

Kann, K., Cotterell, R., and Schütze, H. (2017). Neural multi-source morphological reinflection. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 514–524.

Kirov, C., Cotterell, R., Sylak-Glassman, J., Walther, G., Vylomova, E., Xia, P., Faruqui, M., Mielke, S. J., McCarthy, A. D., Kübler, S., et al. (2018). Unimorph 2.0: Universal morphology. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Kong, L., Dyer, C., and Smith, N. A. (2015). Segmental recurrent neural networks. *arXiv preprint arXiv:1511.06018*.

Littell, P., Pine, A., and Davis, H. (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.

Littell, P. (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, August. Association for Computational Linguistics.

Liu, L. and Hulden, M. (2020). Analogy models for neural word inflection. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2861–2878.

Liu, L. and Hulden, M. (2021a). Backtranslation in neural morphological inflection. In *Proceedings of the Second Workshop on Insights from Negative Results in NLP*, pages 81–88, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Liu, L. and Hulden, M. (2021b). Can a transformer pass the wug test? tuning copying bias in neural morphological inflection models. *arXiv preprint arXiv:2104.06483*.

Max Planck Institute and University of Leipzig. (2008). Leipzig glossing rules.

McCarthy, A. D., Kirov, C., Grella, M., Nidhi, A., Xia, P., Gorman, K., Vylomova, E., Mielke, S. J., Nicolai, G., Silfverberg, M., et al. (2020). Unimorph 3.0: Universal morphology. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3922–3931.

Moeller, S., Kazeminejad, G., Cowell, A., and Hulden, M. (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.

Moeller, S., Liu, L., Yang, C., Kann, K., and Hulden, M. (2020). IGT2P: From interlinear glossed texts to paradigms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5251–5262, Online, November. Association for Computational Linguistics.

Nivre, J., De Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajic, J., Manning, C. D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., et al. (2016). Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666.

Ott, M., Edunov, S., Baevski, A., Fan, A., Gross, S., Ng, N., Grangier, D., and Auli, M. (2019). fairseq: A fast, extensible toolkit for sequence modeling. In *NAACL*.

Rigsby, B. (1986). Gitxsan grammar. Ms., University of Queensland, Australia.

Rigsby, B. (1989). A later view of Gitksan syntax. In M. Key et al., editors, *General and Amerindian Ethnolinguistics: In remembrance of Stanley Newman*. Mouton de Gruyter, Berlin.

Schwartz, L., Chen, E., Hunt, B., and Schreiner, S. L. (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.

Sennrich, R., Haddow, B., and Birch, A. (2016). Improving neural machine translation models with monolingual data. In *54th Annual Meeting of the Association for Computational Linguistics*, pages

86–96. Association for Computational Linguistics (ACL).

Silfverberg, M. and Hulden, M. (2018). An encoder-decoder approach to the paradigm cell filling problem. In *EMNLP*.

Snoek, C., Thunder, D., Loo, K., Arppe, A., Lachler, J., Moshagen, S., and Trosterud, T. (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.

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

Tarpent, M.-L. (1987). *A Grammar of the Nisgha Language*. Ph.D. thesis, University of Victoria.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Wiseman, S. and Rush, A. M. (2016). Sequence-to-sequence learning as beam-search optimization. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1296–1306, Austin, Texas, November. Association for Computational Linguistics.

Wu, S., Cotterell, R., and Hulden, M. (2021). Applying the transformer to character-level transduction. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1901–1907, Online, April. Association for Computational Linguistics.

A. Sample inflection table

A Gitksan inflection table for *'wa* ('to find, reach') generated from IGT and displayed in TSV format. Many cells in the table are empty since they were unattested in the IGT data.

```
ROOT find 'wa 'wa 'wa
ROOT-SX _ _ _ _
ROOT-PL _ _ _ _
ROOT-3PL _ _ _ _
ROOT-ATTR _ _ _ _
ROOT-3.II find-3.II 'wa-t 'wat 'wa-3.II
ROOT-PL-SX _ _ _ _
ROOT-1SG.II _ _ _ _
ROOT-2SG.II _ _ _ _
ROOT-2PL.II _ _ _ _
ROOT-3PL.II find-3PL.II 'wa-diit 'wadiit 'wa-3PL.II
ROOT-1PL.II _ _ _ _
ROOT-PL-3PL _ _ _ _
ROOT-TR-3.II find-TR-3.II 'wa-i-t 'wayit 'wa-TR-3.II
ROOT-PL-3.II _ _ _ _
ROOT-PL-ATTR _ _ _ _
ROOT-PL-2SG.II _ _ _ _
ROOT-TR-1SG.II _ _ _ _
ROOT-PL-3PL.II _ _ _ _
ROOT-PL-1SG.II _ _ _ _
ROOT-TR-1PL.II find-TR-1PL.II 'wa-i-'m 'wayi'm 'wa-TR-1PL.II
ROOT-PL-1PL.II _ _ _ _
ROOT-TR-2PL.II _ _ _ _
ROOT-TR-3PL.II _ _ _ _
ROOT-TR-2SG.II _ _ _ _
ROOT-PL-TR-3.II _ _ _ _
ROOT-PL-TR-2SG.II _ _ _ _
ROOT-PL-TR-3PL.II _ _ _ _
ROOT-PL-TR-1SG.II _ _ _ _
ROOT-PL-TR-1PL.II _ _ _ _
ROOT-PL-TR-2PL.II _ _ _ _
```

B. Inflection table abbreviations

The abbreviations used in the inflection tables are described here with their meaning.

Tag	Description
1/2/3	First/second/third-person agreement. (May combine with SG/PL.)
#SG, #PL	Singular/plural-animate argument agreement. (Combines with 1/2/3.)
ATTR	Attributive marker, creates a modifier.
II	Series II suffixal agreement paradigm for absolutives/ergatives/possessors.
PL	Plural, or iterative pluractional.
SX	Extracted (relativized or focused) intransitive subject.
TR	Transitive marker in independent clauses; ergative agreement follows.