An Inflectional Database for Gitksan

Bruce Oliver
Edith Coates
Clarissa Forbes
Garrett Nicolai
Changbing Yang
Miikka Silfverberg
Farhan Samir
email: msilfver@mail.ubc.ca

Social Sciences and Humanities Research Council of Canada
Canada

Introduction

We present an inflectional database for Gitksan; an endangered Tsimshianic language spoken in British Columbia, Canada.

Our starting-point is interlinear glossed text (IGT) generated in language documentation efforts

We convert the IGT data into a paradigm-level inflectional resource using a combination of rule-based methods and neural morphological reinflection

Multitude of applications: language learning resources, experiments in computational morphology & training of glossing models

Tahltan
Se
Gitxsan
Nisga'a
Nass River
PRINCE RUPERT
Skeena River
Wet'suwet'en
Tsimshian
Haida
Haisla
Heiltsuk
Nuxalk

Gitksan IGT Data

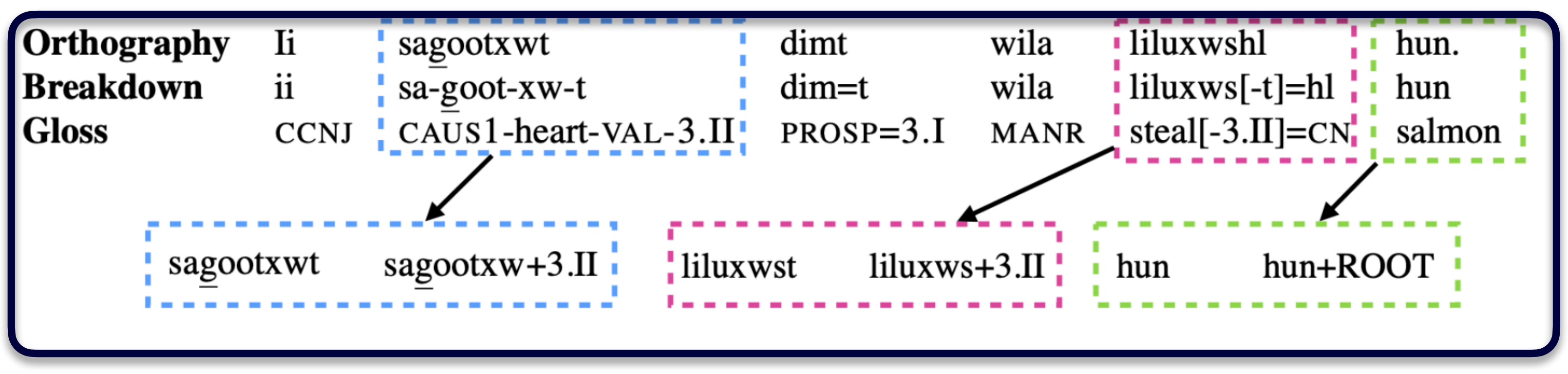
Orthography	Ii	sag̲ootxwt	dimt	wila	liluxwshl	hun.
Segmentation	ii	sa-g̲oot-xw-t	dim=t	wila	liluxws[-t]=hl	hun
Gloss	CCNJ	CAUS1-heart-VAL-3.II	PROSP=3.I	MANR	steal[-3.II]=CN	salmon
Translation	And he planned to steal a fish.					

IGT data sourced from a corpus of glossed Gitksan narratives (Forbes et al., 2017)

The oral narratives come from three speakers from three distinct dialect areas, in roughly equal amounts: Ansbayaxw (Eastern); Gijigyukwhla'a and Git-anyaaw (Western).

Transcribed in the accepted community orthography (Hindle and Rigsby, 1973)

Conversion pipeline



Orthography	Ii	sag̲ootxwt	dimt	wila	liluxwshl	hun.
Breakdown	ii	sa-g̲oot-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

sag̲ootxwt   sag̲ootxw+3.II

liluxwst   liluxws+3.II

hun   hun+ROOT

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
(3) recovery of hidden morphemes like [-t]

We also add +ROOT tags which signify the base form of a lexeme

We exclude words like ii 'and' belonging to closed classes

Paradigm cell-filling experiments

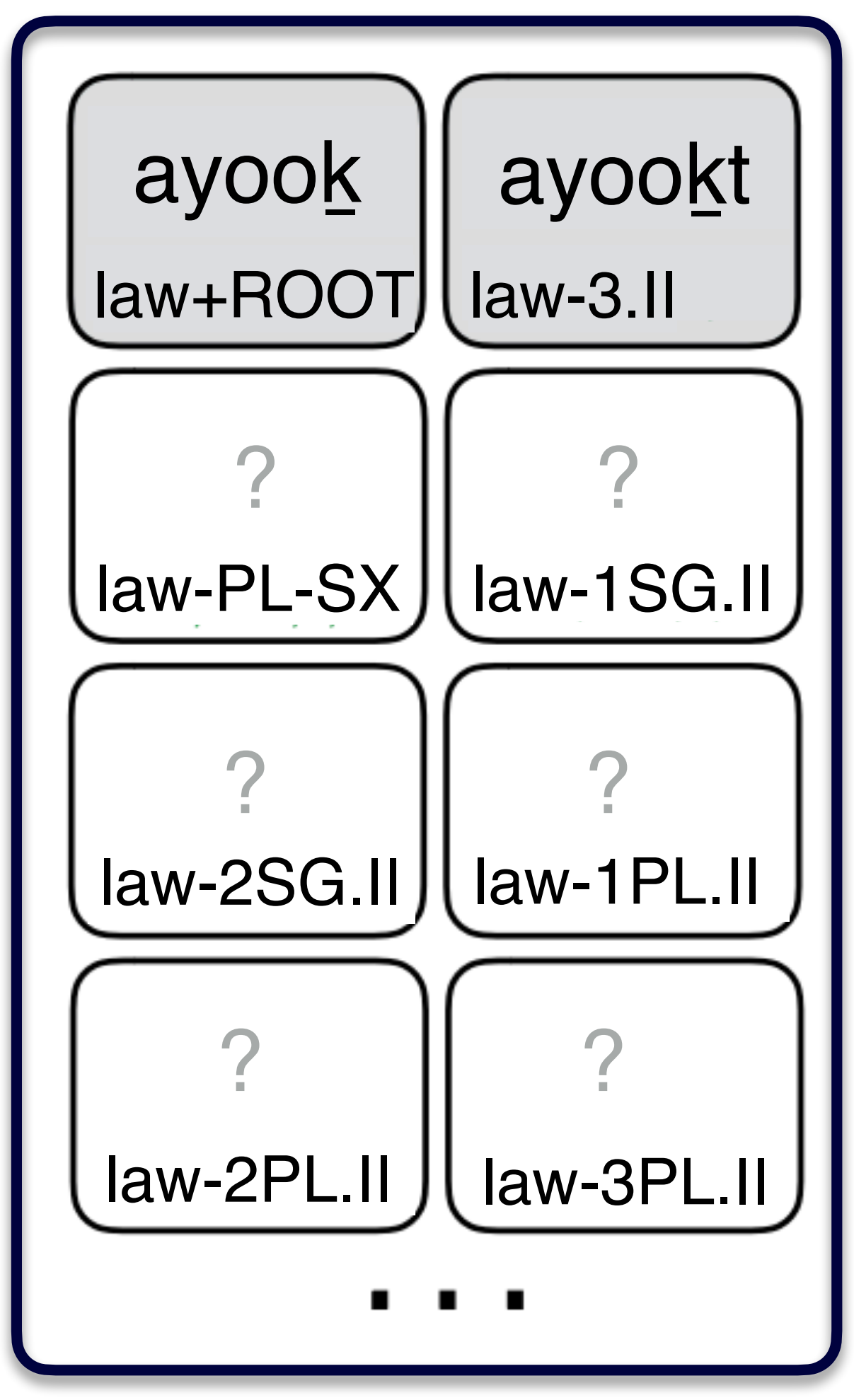
We end up with 1055 inflection tables (containing 31 slots each). In total, there are 2125 inflected forms in the dataset

Only 5.3% of the slots in our tables are filled and 51% of tables contain a single filled slot

Due to the sparsity of the tables, we explore methods to automatically fill in empty slots

We train morphological reinflection models to predict missing slots in the inflection tables, thereby completing partial tables.

We train Fairseq transformers (Ott et al., 2019): 4 layers, 4 heads, 256-dim embeddings, hidden layer size 1024



ayook̲ law+ROOT	ayook̲t law-3.II
? law-PL-SX	? law-1SG.II
? law-2SG.II	? law-1PL.II
? law-2PL.II	? law-3PL.II

. . .

Generating data for experiments

We split our data into three disjoint sets:
(1) a train set (858 forms),
(2) a dev set (302 forms)
(3) a test set (124 forms)

None of the lexemes in the test set are observed during training

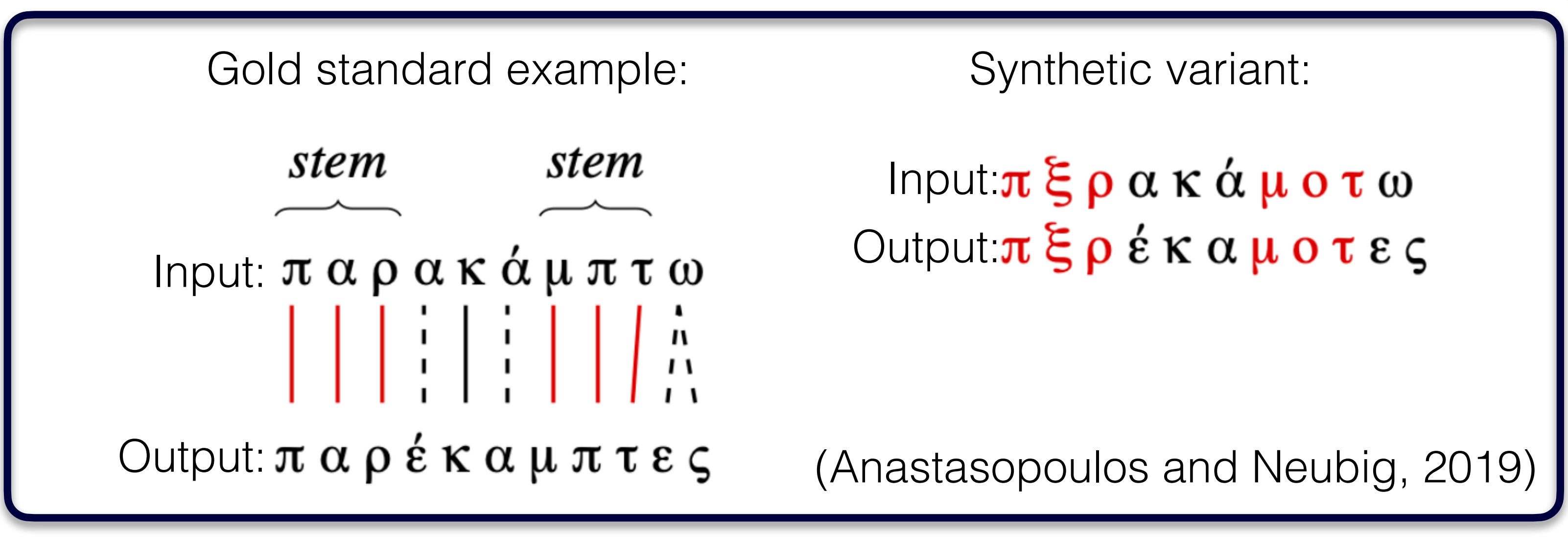
We frame form prediction as reinflection and train the model on our train set:

```
niye'et    3.II          n i y e ' e t IN(3.II) OUT(ROOT)         n i y e ' e
niye'e     ROOT    ➡     n i y e ' e t IN(3.II) OUT(1SG.II)       n i y e ' e ' y
niye'e'y   1SG.II        n i y e ' e IN(ROOT) OUT(3.II)           n i y e ' e t
                         n i y e ' e IN(ROOT) OUT(1SG.II)         n i y e ' e ' y
                         n i y e ' e ' y IN(1SG.II) OUT(ROOT)     n i y e ' e
                         n i y e ' e ' y IN(1SG.II) OUT(3.II)     n i y e ' e t
```

During test time, 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. We apply either majority voting or choose one of the outputs randomly (baseline)

Data augmentation


Gold standard example:
stem
stem
Input: π α ρ α κ ά μ π τ ω
Output: π α ρ έ κ α μ π τ ε ς
Synthetic variant:
Input:π ξ ρ α κ ά μ ο τ ω
Output:π ξ ρ έ κ α μ ο τ ε ς

(Anastasopoulos and Neubig, 2019)

We experiment with two data augmentation strategies:

(1) Data hallucination, where synthetic examples are generated by introducing noise into existing gold standard training examples (Anastasopoulos and Neubig, 2019)
(2) Back-translation, where self-training on predicted forms is used to boost model performance (Sennrich et al., 2016; Liu et Hulden, 2021)

Results

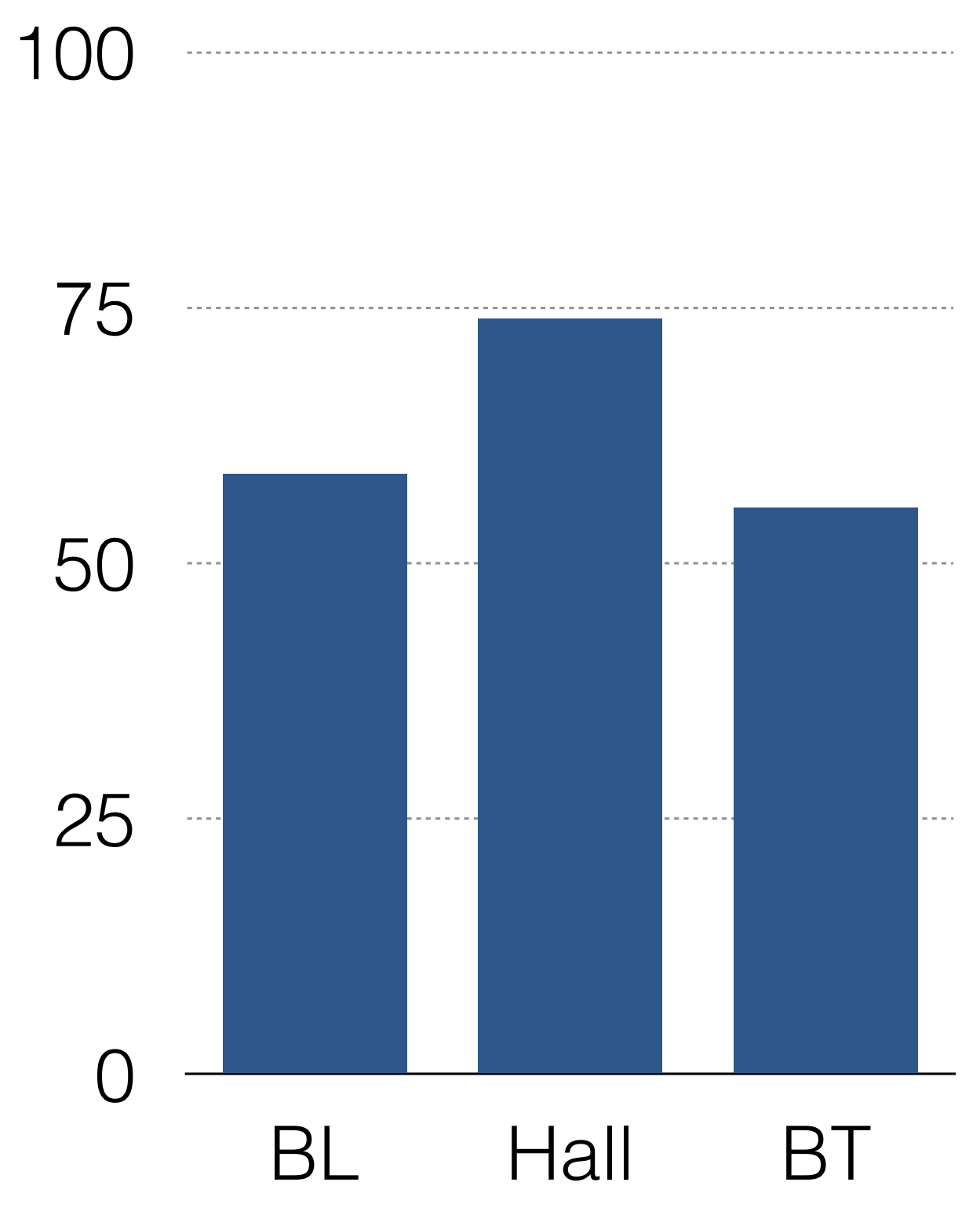
Data hallucination provides the best results, 74% full-form accuracy

This presents an improvement of roughly 15%-points over the baseline system.

Back-translation does not result in improvements over the baseline

Erroneous forms typically have edit distance 1-2 from the correct form

Manual verification and correction will be required, but the effort will likely be moderate



BL - baseline
Hall - data hallucination
BT - back-translation

Discussion

Some manual post-processing will be required due to errors in paradigm cell-filling

Data augmentation is crucial in improving reinflection performance!

We have contributed and expanded a new data resource but the work is not yet complete

The IGT itself is undergoing revision, and dictionary resources which include part of speech information are under construction.

Our inflection tables will be regenerated to account for this changes

References:
Anastasopoulos, A. and Neubig, G. 2019. Pushing the limits of low-resource morphological inflection. *EMNLP*
Forbes, C., Nicolai, G., and Silfverberg, M. 2021. An FST morphological analyzer for the Gitksan language. *SIGMORPHON*
Hindle, L. and Rigsby, B. 1973. A short practical dictionary of the Gitksan language. *Northwest Anthropological Research Notes.*
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. *NAACL*
Sennrich, R., Haddow, B., and Birch, A. 2016. Improving neural machine translation models with monolingual data. *ACL*