Dim Wihl Gat Tun: The Case for Linguistic Expertise in NLP for Underdocumented Languages

Clarissa Forbes[3] **Farhan Samir**[ʕ] **Bruce Harold Oliver**[ʕ] **Changbing Yang**[ʕ]
Edith Coates[ʕ] **Garrett Nicolai**[ʕ] **Miikka Silfverberg**[ʕ]
[3]Independent Researcher [ʕ]University of British Columbia
[3]forbesc@alumni.ubc.ca [ʕ]first.last@ubc.ca

Abstract

Recent progress in NLP is driven by pretrained models leveraging massive datasets and has predominantly benefited the world's political and economic superpowers. Technologically underserved languages are left behind because they lack such resources. Hundreds of underserved languages, nevertheless, have available data sources in the form of interlinear glossed text (IGT) from language documentation efforts. IGT remains underutilized in NLP work, perhaps because its annotations are only semi-structured and often language-specific. With this paper, we make the case that IGT data can be leveraged successfully provided that target language expertise is available. We specifically advocate for collaboration with documentary linguists. Our paper provides a roadmap for successful projects utilizing IGT data: (1) It is essential to define which NLP tasks can be accomplished with the given IGT data and how these will benefit the speech community. (2) Great care and target language expertise is required when converting the data into structured formats commonly employed in NLP. (3) Task-specific and user-specific evaluation can help to ascertain that the tools which are created benefit the target language speech community. We illustrate each step through a case study on developing a morphological reinflection system for the Tsimchianic language Gitksan.

1 Introduction

Progress[1][2] in NLP research has primarily manifested in tools for the world's political and economic superpowers (Blasi et al., 2021), and it is unclear how we can build more inclusive language technologies. Even multilingual pretraining methods (e.g., Liu et al., 2020; Artetxe et al., 2018), capable of producing effective models in the absence of large annotated training datasets require unannotated corpora that are prohibitively large for 90% of the world's languages (Joshi et al., 2020).

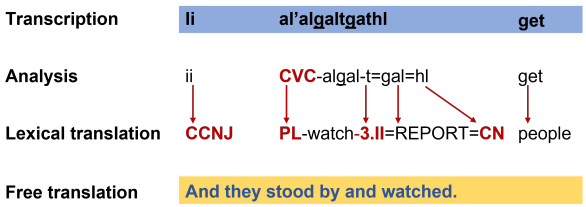

Figure 1: An example of Gitksan interlinear glossed text (IGT). The text contains four levels of annotation: (1) An orthographic transcription, (2) A segmentation into normalized component morphemes (CVC refers to the reduplicated segment *al'*), (3) an interlinear gloss and (4) an English translation.

Nevertheless, many languages in this 90% have a body of resources. Language documentation and linguistic fieldwork are an ongoing task worldwide, and many resources continue to be developed in these traditions (Bird, 2020). We have access to wordlists, bilingual dictionaries for over 1000 languages (Wu et al., 2020), aligned speech recordings for over 700 languages (Black, 2019), multi-parallel texts for 1600+ languages (McCarthy et al., 2020b), and knowledge of related languages (Haspelmath et al., 2005). Indeed, researchers have leveraged these resources to build impressive, useful computational systems for multilingual morphological analyzers (Nicolai and Yarowsky, 2019), adapting pretrained language models for over 1000 languages (Ebrahimi and Kann, 2021), and building massively multilingual speech recognition systems (Adams et al., 2019), among others.

There are additional language documentation resources which have yet to be fully leveraged in the aim to produce more inclusive language technology. Interlinear glossed texts (IGTs) depicted in Figure 1 are semi-structured texts which comprise not only monolingual corpus data (e.g. *al'algaltgathl*) but also morpheme-level segmentations (e.g. *CVC~algal-t=gat=hl*), glosses for component-morphemes (e.g. *PL~watch-*

[1]*Dim wihl gat tun* - "This is what the people should do"
[2]First two authors contributed equally.

Findings of the Association for Computational Linguistics: ACL 2022, pages 2116 - 2130
May 22-27, 2022 ©2022 Association for Computational Linguistics

3.II=REPORT=CN), word alignment information (Zhao et al., 2020), and free translations. IGTs remain a major annotated datatype produced in the course of linguistic fieldwork: examples are continuously digitized in large databases for hundreds of languages (Lewis and Xia, 2010), and entire corpora of IGT are periodically published in volume series such as *Texts in Indigenous Languages of the Americas*. They have the potential to serve as training data for a wide variety of computational systems including bilingual lexicons, morphological analyzers, dependency parsers, part-of-speech taggers, and word-aligners (Georgi, 2014). Yet while they are accessible, they remain severely underutilized for these purposes.

Part of the general hesitancy in adoption of IGT as training data may lie in the fact that the annotation format is only semi-structured and often language-specific. While the general IGT format is governed by the Leipzig glossing rules (Comrie et al., 2015), there remains significant flexibility for the annotator to customize tags and conventions for any given language. This makes IGT challenging as a format for training supervised NLP models.

With this paper, we make the case that IGT data can be leveraged in NLP research and language applications for speech communities, provided that target language expertise is available. Specifically, we argue that it is essential to collaborate with documentary linguists who are familiar with the language-specific annotations in the IGT data in order to leverage the data for NLP tasks. This may furthermore provide a foundation for co-designing language technologies with a given speech community (Bird, 2020).

Our paper provides a roadmap, portrayed in Fig. 2, for navigating three areas of significant uncertainty that arise when incorporating IGT data for inclusive language technology. First, we need to define what NLP tasks can be accomplished with a given set of IGT data, and whether they are of value to the speech community. Second, after selecting useful tasks, we will need to preprocess the data, potentially by converting it to a structured format commonly employed in NLP tasks. Finally, we need task-specific and user-specific evaluation procedures in order to be explicit about the failure modes of the technology, as it is ultimately being developed for end users like speakers and linguists rather than solely comparison with other researchers.

We focus on the first two of these areas, forwarding our argument through a case study on developing a morphological reinflection system for the Gitksan language (Section 2.3) that has applications in language teaching.

2 Background

2.1 NLP for Underdocumented Languages

Computational work on underdocumented and low-resource languages has accelerated in recent years due to increasing recognition of both the role of NLP in language preservation as well as dedicated workshops like ComputEL (Arppe et al., 2021), AmericasNLP (Mager et al., 2021) and SIGTYP (Vylomova et al., 2021). Most of this work aims to assist in language documentation and revitalization, with machine translation being another important research area. Mager et al. (2018) and Littell et al. (2018) present surveys of existing NLP tools for the North American Indigenous languages, many of which are underdocumented, and discuss core challenges: morphological complexity, limited training data, and dialectal variation.

Several authors have trained NLP models on IGT to accelerate language documentation, with automatic glossing being a prominent research direction. The first approaches simply memorized earlier glossing decisions and enabled the annotator to re-use these later (Baines, 2009). Later approaches have relied on structured models like CRFs (McMillan-Major, 2020), RNN encoder-decoders (Moeller and Hulden, 2018) and transformers (Zhao et al., 2020) to generate glosses for unseen tokens. NLP techniques can also be used to generate inflection tables from IGT (Moeller et al., 2020). These find applications both in language documentation and language education, often to facilitate the production of more IGT data. A related approach is to generate morphological analyzers using IGT as a starting-point (Zamaraeva, 2016; Wax, 2014).

Several papers discuss challenges related to IGT as a data type. One of the principal concerns is the noisiness of the annotations (Moeller et al., 2020). This problem is compounded by the fact that annotation schemas employed by linguists preparing IGT tend to be idiosyncratic[3] and often lack internal consistency (Baldridge and Palmer, 2009; Palmer et al., 2009). The design of annotation stan-

[3]These systems are well motivated but unlikely to be easily comparable with other annotation schemas.

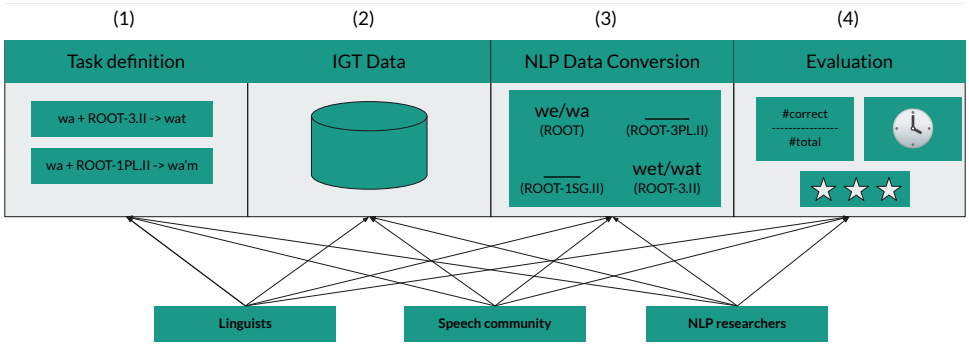

Figure 2: A roadmap for incorporating Interlinear Glossed Text (IGT) data for building more inclusive language technology. (1) We first need to define what NLP tasks can be accomplished with a given set of IGT data and whether they are valuable to the speech community (see Section 2.3). (2) Next, we need to gather the relevant IGT data that was created during linguistic fieldwork with the speech community (see Section 2.3). (3) Next, the IGT data needs to be converted to a structured format amenable for NLP formats. (4) The model needs to be evaluated not only in terms of standard NLP model selection metrics but also for efficacy for end-users such as efficiency in time-savings and usability (see Section 4). Crucially, all three stakeholders – speech community members, NLP researchers, and linguists – should be involved throughout the process.

dards is important: Zhao et al. (2020) note that this can have an impact on the performance of glossing systems. McMillan-Major (2020) notes a further challenge: IGT often includes not only morphological information, but also syntactic, semantic, and pragmatic annotations, which can be much harder to learn in low-resource settings.

In addition to challenges in the IGT data type itself, there are other challenges in NLP applications for underdocumented languages. Ward and Genabith (2003) discuss many problems related to development of computer-assisted language learning for endangered languages: lack of orthographic standards, limited resources, and limited documentation of the language. van Esch et al. (2019) also discuss NLP tools that can be helpful for documentation of low-resource languages, but they note that restrictive licenses can often be problematic for engineering.

2.2 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). 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 wordlists and workbooks, but no comprehensive dictionary, grammar, or pedagogical curriculum. There is an accepted orthography (Hindle and Rigsby, 1973), and a talking dictionary mobile app in active use by the community (*Mother Tongues Dictionaries*, formerly *Waldayu*; Littell et al. (2017)).

Other computational studies interact with the active documentation efforts surrounding Gitksan to produce new frameworks and resources. Dunham et al. (2014) present a database structure for hosting audio and transcribed data in language documentation contexts, adopted for Gitksan and eight other underdocumented languages. Littell et al. (2017) present a dictionary interface which is capable of fuzzy search. They mention this specifically as a way to increase accessibility in a setting where orthographies have not been standardized or where many users are language learners. Forbes et al. (2021) present a finite-state morphological

analyzer for Gitksan; they test coverage across different dialects of Gitksan and use handcrafted rules to increase coverage for spelling variants.

2.3 Constructing a Gitksan Pedagogical Application from IGT Data

Our project generates language learning exercises for Gitksan grammar. The need for these exercises was identified in discussions with documentary linguists working on Gitksan (the task definition step in Figure 2). Specifically, our goal is to automatically generate exercises for noun and verb inflection. As source material, we use Gitksan IGT data collected by linguists at the University of British Columbia for language documentation purposes (the data step in Figure 2). Examples of this data are shown in Figure 1 and Appendix A.

Due to extensive morphological annotation, IGT provides a valuable starting point for our work. However, the annotations are far too detailed for our purposes—many derivational affixes are annotated in the data (further discussed in Section 3.1). These are irrelevant and can be downright harmful for grammar exercises. To remedy this misalignment between the raw IGT data and our NLP task, we collaborate with Gitksan documentary linguists to identify a set of inflected forms with clearly defined grammatical function, while discarding derivational morphology. We then convert the IGT data into a set of inflectional paradigms (the data conversion step in Figure 2). We further discuss this conversion process in Sections 3.2 and 3.3. Since the inflectional paradigms sourced from corpora are sparse,[4] we train models to fill in missing forms (Section 4). This is more widely know as the Paradigm Cell-Filling Problem (PCFP) (e.g., Silfverberg and Hulden, 2018). We then evaluate the system on it's capacity to automatically generate inflections, and discuss limitations of our current evaluation procedure (the evaluation step in Figure 2).

3 Challenges in Incorporating IGT into NLP Research

Because tokens in IGT are already segmented and annotated, it forms an ostensibly convenient starting-point for further processing and token-based grouping. In many ways, IGT is, however, a challenging data type for use in pedagogical and NLP applications. This section presents three specific challenges posed by IGT data when NLP techniques are applied. First, while IGT will contain a wealth of useful information for NLP models, it might also contain information which is far too fine-grained for automatic learning purposes, at least given the quantity of data which are available. Second, IGT often contain idiosyncratic or language-specific conventions which may not be easily converted to or represented in standardized frameworks. Third, because IGT is used as a device for language documentation, it will often contain dialectal variation, an important meta-characteristic which in aggregate cannot be easily distinguished from other types of variation or spelling errors. We argue that handling these issues for successful data preprocessing requires consultation with linguistic experts, and exemplify with instances from the Gitksan IGT and our use-case.

3.1 Annotation Granularity

Documentary linguists' goals when annotating IGT is to present an accurate representation of the surface phonology and morphology of a given utterance, as well as the syntactic and semantic information contributed by its component morphemes, with fine attention to detail given the rarity and value of the data. This goal of providing fine-grained annotations and transcriptions, however, can be in conflict with the NLP research aim of building models that can generalize in the real world (i.e., future elicited linguistic data). The fine-grained details are often extraneous for the purposes of building NLP models, and can counterproductively act as noise that makes learning systematic patterns more difficult.

As an example of this mismatch in disciplinary goals, consider the sample IGT token in (1).

(1) `maaxwsxwa`
`maaxws-xw-a`
`fallen.snow-VAL-ATTR`
'white'

In this token, the productive stem is deconstructed into a historical root (*maaxws*) and a derivational suffix (*-xw*)—along with an inflectional affix (*-a*). It is unclear from the input that the most readily recognizable lexical stem in this form is the larger unit *maaxwsxw* 'white, snow-colored', and that the internal boundaries within that stem reference etymological and derivational information not relevant to the typical NLP task. The derivational and

[4] Due to the Zipfian distribution of language (Blevins et al., 2017).

inflectional affixes are not differentiated in IGT.[5]

At first glance, it might seem reasonable to train an NLP model to automatically generate such a gloss for Gitksan input words in an effort to accelerate language documentation. While this remains one of the most common NLP tasks associated with IGT, it may be difficult for models to deliver high performance if the IGT input, like Gitksan's, contains a substantial proportion of derivational and etymological information, since this information is lexical and unpredictable.

Collaboration with documentary linguists, in addition to being important when a project aims to improve the documentary linguistic workflow, can be useful for identifying these aspects of the data which may be less valuable to learn. This information can be applied in data preprocessing to improve model performance given data scarcity. For the token in (1), an alternative segmentation *maaxwsxw-a* into a word stem and a productive inflectional affix *white-ATTR* is more amenable to both automated labeling and inflection tasks, particularly in low-resource conditions. Furthermore, reference to derivational information is unnecessary in our use case of performing automated inflection for use in a pedagogical application. We collaborated with documentary linguists familiar with Gitksan to manually filter morphology into derivational versus inflectional, to determine whether an affix should be classed as part of a lexical stem or should signal a paradigm cell in the inflectional template. This allowed derivational morphology to be effectively excluded before we moved to the paradigm cell-filling task. This filtering process was non-trivial, requiring solid understanding of the target language, its description, and its vocabulary.

3.2 Using Existing Annotation Standards

The annotation schemas employed in IGT are often idiosyncratic (Palmer et al., 2009; Comrie et al., 2015), which typically makes them better suited for language documentation than NLP tasks. When aiming to leverage IGT data for use in NLP tasks, we must then consider on a case-by-case basis whether it is more beneficial to convert the IGT data to an NLP-standard format, or work with the IGT annotations largely as-is, adapting them to our specific needs. Relevant to this decision are factors such as how labor-intensive the conversion will be, how well the standard format accommodates linguistic information that has been detailed in the IGT, and whether conversion of the dataset to the standard format aligns with specific project goals and speech community interests.

The possible format that we consider for annotating inflection tables is the Unimorph standard (McCarthy et al., 2020a; Sylak-Glassman, 2016), a popular schema for annotation of inflectional morphology that can facilitate cross-lingual transfer by enabling language-independent annotations. Ultimately, we opted to adapt the Gitksan IGT to our specific needs after determining that conversion would be extremely labor-intensive, and that several types of information in the Gitksan IGT could not be represented in the UniMorph standard. We present three of the most significant issues:

1. Part-of-Speech The Unimorph standard relies on part-of-speech (POS) tags as a major component of word form annotation. However, POS information is frequently not annotated in IGT (Moeller et al., 2020), and no POS information was included in our Gitksan IGT.

For some underdocumented languages, POS information requires substantial experience and manual attention to annotate. For example, our target language Gitksan displays considerable category flexibility, meaning that syntactic and morphological behavior can cross word class boundaries. In Gitksan, the inflectional paradigms of nouns and verbs overlap substantially. As an example, agreement markers can affix to both nouns and verbs, conveying a number of functions. Some are exemplified in (2). As a consequence, in Gitksan it is difficult to use morphological inflection to deduce a lexeme's POS.

(2) Forms with *-'y* (1SG series II)

a. *hlguuhlxwi**'y*** - my child (POSSR)
b. *yee**'y*** - I walked (ABS)
c. *t'agi**'y*** - x forgot me (ABS, dependent)
d. *t'agi**'y*** - I forgot x (ERG)

In addition, Gitksan nouns and verbs are syntactically flexible, meaning that Gitksan nouns can function as verbs in text, and vice versa. For example, a noun *g̲anaa'w* 'frog' can be used predicatively without a copula in main verb position in the sentence *Hlaa ap **g̲anaa'wi'y*** 'I'm a frog now'. It takes absolutive inflection when it does so. Due

[5] For an English analogue, consider splitting the lexicalized verb *enforce* into a prefix *en-* and root *force*. The *en-* prefix is recognizable, but not productive or relevant to inflection tasks.

to this morphological and syntactic flexibility, a 1SG-inflected noun like *g̲anaa'wi'y* could be annotated two ways in UniMorph depending on the context (`frog;PSS1S` versus `frog;1SG;ABS`[6])—yet in the IGT, they are uniformly annotated as `frog-1SG.II`. Reviewing the contextual function of every noun and verb in the IGT dataset to apply the appropriate UniMorph tags would require an infeasible amount of expert reannotation.

2. Inflection vs. derivation Unimorph postulates a strict division into inflectional and derivational morphology (and only annotates inflectional morphology). The IGT format has no such division, because it can be used to represent morphology at any level of granularity the annotator wishes.

We have mentioned in Section 3.1 that determining the difference between inflectional and derivational morphology from IGT input is non-trivial. For example, the Gitksan morpheme *-xw* has a variety of uses which might be considered more derivation-like (D) or more inflection-like (I).

- Creating intransitive predicates from nouns: *os**xw*** 'have a dog' from *os* 'dog' (D)
- Marking inchoatives: *mit**xw*** 'be full' vs. causative *midin* 'fill' (D)
- Marking passives: *jap**xw*** 'be made' from transitive *jap* 'do, make' (D?)
- Marking verbs with certain preverbs: *sik'ihl huut**xw*** 'try to run away' vs. *huut* 'run away' (I?)
- Optional in some possessives: *la̲xyip**xw**si'm* 'your.pl land' vs. *la̲xyipsi'm* 'your.pl land' (?)

This morpheme's uses and degree of productivity are still little-understood, so its status as inflectional or derivational remains unclear.[7] For now, we provisionally exclude this morpheme from our inflection tables as 'derivational'. In a UniMorph system, this morpheme's exclusion or inclusion in the annotation would constitute a prematurely strong claim about whether it was inflectional, and the tagset used to annotate it likewise a prematurely strong claim about its function.

3. Clitics Gitksan is rich in clitics, annotated with the equals sign in IGT '='. Their attachment is determined by prosodic and linear factors. Prenominal clitics are illustrated in example (3).

(3)
```
Giigwis        Maryhl  g̲ayt.
giikw-i[-t]=s  Mary=hl g̲ayt
buy-TR-3.II=PN Mary=CN hat
```
'Mary bought a hat.'

In the example above, the proper noun clitic *=s* attaches to the verb but is syntactically associated with *Mary*. The common noun clitic *=hl* attaches to *Mary* but is associated with *g̲ayt* 'hat'. Since UniMorph does not annotate such cross-token dependencies (or other clitics), this central feature of Gitksan cannot be represented.

Recommendations Current computational morphology research relies heavily on standardized tagsets like UniMorph, in particular for crosslingual transfer (Anastasopoulos and Neubig, 2019). However, these formats can be either labor-intensive or impossible to apply to underdocumented language datasets, depending on the idiosyncratic conventions of a given IGT and language-specific factors. Our understanding of the language may not be sufficiently mature to implement some of UniMorph's strict requirements, or important phenomena may fall outside of the defined scope of UniMorph. We recommend that NLP projects on underdocumented languages collaborate with language experts to determine where language-agnostic data formats can be applied, and to design project-specific data formats as needed.

3.3 Dialectal variation

Dialectal variation is a pervasive feature of languages worldwide, from English (consider African-American English and Standard American English; Blodgett et al., 2016) to Arabic (consider Modern Standard Arabic and the Doha dialect; Kumar et al., 2021). Many Indigenous languages of North America also exhibit vast dialectal variety, with significant variance in the level of mutual intelligibility between languages and dialects (Mithun, 2001, Ch.6).

Although Gitksan has an estimated fewer than 1K speakers, each village has a different way of speaking, and the speech community recognizes two salient dialects (Eastern/Upriver and Western/Downriver). Gitksan dialectal variation is typically reflected in written materials due to the lack of a widely-adopted orthographic standard which

[6] Other clause type features would be required here but it remains unclear how best to represent Gitksan's clause-typing system with UniMorph labels.

[7] Elsewhere, some linguistic descriptions present cases of morphology which do not fit into conventional delineations of the inflectional/derivational divide, such as plural/pluractional markers in Halkomelem Salish (Wiltschko, 2008).

would 'flatten' it.[8] For many underdocumented languages, written orthographies have been in use for a relatively short period of time, and communities place different levels of emphasis on literacy and standardization versus conversational fluency. As a consequence, orthographic conventions can vary widely across dialects and writers in low-resource and underdocumented language contexts.

It is desirable in building inclusive language technology to accommodate and reflect variation, rather than aim to model a homogenous standard form of the language. In building pedagogical resources for language revitalization, we furthermore need to mindfully consider potential data biases as well as what kinds of variation are presented to the user, to avoid implictly suggesting that certain dialects favored for preservation and teaching, which risks reinforcing or creating negative social hierarchies (Demszky et al., 2021).

The first step to ensuring dialectal fairness and appropriate handling of variation in NLP applications is to understand what types of variation are at play, and in particular what dialect a given token belongs to. This allows us to proactively control what data is presented to a user and, for example, ensure that data from different dialects is not mixed together inappropriately. This task is non-trivial: expertise in the language is crucial in order to determine what types of variation are dialectal, and which are idiosyncratic or purely orthographic, including typos and spelling errors. As an example from Gitksan, *gat* and *get* are highly salient East/West dialect variants, while *hun* and *hon* are less-salient variants within the Eastern dialect; *amxsiwaa* and *amxsiiwaa* are two non-dialectal variants of the same word (spelling error/variant), while *sipxw* and *siipxw* are different lexemes.[9] Presently, we include all lexeme variants as separate entries in our inflection tables, enabling us to represent all dialects during training.

Recommendations Distinguishing between different types of variation in the source material is a challenging task but also a crucial one. Expertise in the target language and dialects is required for classifying types of variation, and so language experts are a vital asset for this process. Documentary linguists or community members may have direct information about the dialectal background of speakers that are represented in the data, which is useful for modeling, and will likely have information about how dialectal variation is viewed in the speech community (e.g. it may be highly politicized), which is important for application design.

Variation is not only an important issue when constructing datasets. It is also essential to evaluate the final model's performance according to the principle of dialectal fairness (Choudhury and Deshpande, 2021) Recently, measures for dialect fairness have emerged in the NLP community: Faisal et al. (2021) and Kumar et al. (2021) advocate for computing performance separately for each dialect rather than computing a single macro average performance figure over distinct dialects. They also propose to use standard deviation between system performance on different dialects and the generalized entropy index (Speicher et al., 2018) as measures for dialectal unfairness which we naturally want to minimize.

4 Steps toward Building a Language Learning Application

The inflectional paradigms collected from the adapted IGT corpus are overly sparse for automatically generating pedagogical exercises. To automatically fill in these paradigms, an example of which is shown in Appendix B, we train and evaluate a morphological reinflection system.[10]

Data We train and test reinflection models on the Gitksan morphological paradigms described in Section 3. We generate three splits of the data from our complete set of paradigms: train ($N = 858$ word forms), validation ($N = 302$ word forms), and test ($N = 124$ word forms) data splits.

Training We form training pairs by using the given forms in each table and learn to reinflect each given form in a table to another given form in the same table, following Silfverberg and Hulden (2018). Model parameters are shown in Appendix C.

[8] Linguistic description frequently aims to record dialectal and even speaker-level variation. Our datasets are based on IGT data which explicitly annotates such variation in the orthographic representation.

[9] In IGT the gloss cannot always be used to differentiate lexemes. Depending on the convention, the same lexeme may appear with different glosses in different contexts (e.g. *'wa*: 'find' or 'reach'), and different lexemes may have the same gloss (e.g. *yook̲* and *gup*: 'eat', which differ on other grounds – transitivity). The latter forms which share a gloss must also be differentiated as lexical variants, not dialectal variants.

[10] Code and data for this experiment is available at https://github.com/smfsamir/gitksan-data.

Evaluation During test time, we predict forms for missing slots based on each of the given forms in the table and take a majority vote of the predictions. We evaluate accuracy on the test set by counting the number of the 124 forms that were correctly predicted. We find that the Transformer model generates 87.09% of the test forms correctly.

Analysis. Our model provides strong performance when measured by the standard metric of accuracy, in particular considering that it is trained on only 858 examples. Accuracy, however, only provides one perspective on the efficacy of the model (Ethayarajh and Jurafsky, 2020). The appropriate evaluation of the system is highly context dependent: For our goal of generating language learning exercises, we want to evaluate whether our system and automatically generated grammar exercises allow for more effective language learning; raw accuracy gleans little insight to the effectiveness of the system for this goal. If in contrast our goal was to facilitate language documentation, we would want to evaluate whether the model gives an overall significant reduction in documentation effort—this largely depends on whether the automatic annotations are of sufficient quality that correcting remaining errors takes less time than annotating all the data from scratch. Further research, in collaboration with documentary linguists and the speech community, is required to determine whether our system can achieve the desired goals of building more practical, inclusive language technology.

5 Discussion

Incorporating IGT data for NLP Language documentation provides a valuable data source for many so called "left-behind" languages (Joshi et al., 2020), which lack traditional annotated and unannotated NLP datasets. For example, IGT data can be used to train systems for morphological inflection, segmentation and automatic glossing, among other applications. Nevertheless, the annotations in IGT are rarely ideally suited for typical NLP tasks, and may need to be significantly adapted. This will typically be hard without extensive knowledge of the target language and annotation conventions which were employed when the IGT data were generated. Linguists and community language experts are well-positioned to address questions related to IGT usability, the structure of the target language, variation in the data, and other annotations in the source data. Collaboration with language experts is not only vital for successful data preprocessing and conversion to the formats required for the typical NLP task, but can also naturally help define research goals and drive the project toward them.

Inclusive Research Goals NLP technologies for underdocumented languages have the capacity to speed up language documentation (e.g., Anastasopoulos, 2019); assisting language revitalization (e.g., Rijhwani et al., 2020; Lane and Bird, 2020); and creating digital infrastructure (e.g., Anastasopoulos and Neubig, 2019). These high-level goals are only a part of what it may mean to create inclusive language technology. Equally valuable as a research goal may be **inclusion**: for speech communities to be acknowledged and engaged in the course of the the research project.[11] We encourage NLP projects on low-resource, minoritized, and/or endangered languages to begin by understanding the speech community context, proceed with community collaboration or endorsement, and ultimately produce concrete benefits that speech communities recognize. This might include outcomes for language teaching and pedagogy, or training opportunities in technology or research.

Evaluation methods can be compiled which address NLP researchers, linguists, and communities' overlapping and divergent goals. For example, pedagogical tools can be directly evaluated for dialect fairness and user/learner improvement.

Practical Collaboration We suggest seeking out opportunities to collaborate directly with community members, in order to solicit their specific expertise when setting the research agenda (i.e. task definition) and conducting evaluation (Czaykowska-Higgins, 2009; Bird, 2020). When the NLP researcher has no existing contact or history with the speech community, this can be pursued via collaboration with a documentary linguist with established community relationships and a similar desire to engage in this research model. Recognize that in any collaboration, different individuals contribute different skills and experience (e.g. pedagogy, annotation, knowledge of community attitudes) and may have different goals and preferred ways of participating, which should simply be discussed within the partnership to ensure things run smoothly.

Research accessibility In discussing inclusive lan-

[11] Underdocumented languages are often the cultural heritage of typically marginalized peoples, sometimes with a history of their data being exploited for political or commercial purposes. NLP research without community involvement may feel like a continuation of this pattern.

guage technologies, we also consider the accessibility of NLP workshops to speech communities, in particular where venues have a dedicated focus on low-resource languages. We note that such venues are often inaccessible to communities due to factors such as the cost of registration. Similarly-oriented workshops in linguistics (e.g. SAIL, WSCLA, family-specific conferences) typically have a tiered registration structure enabling community members to attend for free or minimal cost (e.g. $25). It is worth recognizing that community members are research stakeholders, and ensuring that venues are open to their participation.

6 Conclusion

Although a majority of the world's languages lack the kind of large annotated and massive unannotated datasets which are used to train modern NLP models for high-resourced languages like English (Joshi et al., 2020; Blasi et al., 2021), many languages have other potential data sources such as language documentation data, which so far have remained under-explored. However, care must be taken when applying this type of data, which originally is not intended for NLP use. This is important to ensure that the resulting technologies actually achieve their intended goals like accelerated language documentation or genuinely helpful computer-assisted language learning.

Collaboration with linguists can provide the expertise necessary to engage in modeling with IGT data for underdocumented languages. Linguists can help define an NLP task with good value propositions, given their familiarity and connections with the speech community. They can provide guidance on navigating the IGT format so that we can extract the most useful information for the task at hand. Finally, they can assist in evaluating whether the model achieves appropriate performance on the speech community use cases, and provide feedback on metrics for model success and fairness across dialects. Throughout the development process, documentary linguists and speech community members should be consulted. This will further a greater understanding of the source data and lead to more equitable and effective technologies.

7 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. We also want to thank Samantha Quinto for assisting with visual design. 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

Oliver Adams, Matthew Wiesner, Shinji Watanabe, and David Yarowsky. 2019. Massively multilingual adversarial speech recognition. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Antonios Anastasopoulos. 2019. *Computational Tools for Endangered Language Documentation*. Ph.D. thesis, University Of Notre Dame.

Antonios Anastasopoulos and Graham Neubig. 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*.

Antti Arppe, Jeff Good, Atticus Harrigan, Mans Hulden, Jordan Lachler, Sarah Moeller, Alexis Palmer, Miikka Silfverberg, and Lane Schwartz, editors. 2021. *Proceedings of the 4th Workshop on the Use of Computational Methods in the Study of Endangered Languages*.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. Unsupervised statistical machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

David Baines. 2009. Fieldworks language explorer (flex). *eLEX2009*, page 27.

Jason Baldridge and Alexis Palmer. 2009. How well does active learning actually work? Time-based evaluation of cost-reduction strategies for language documentation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*.

Steven Bird. 2020. Decolonising speech and language technology. In *Proceedings of the 28th International Conference on Computational Linguistics*.

Alan W Black. 2019. Cmu wilderness multilingual speech dataset. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing*.

Damián Blasi, Antonios Anastasopoulos, and Graham Neubig. 2021. Systematic inequalities in language technology performance across the world's languages. *arXiv preprint arXiv:2110.06733*.

James P Blevins, Petar Milin, and Michael Ramscar. 2017. The zipfian paradigm cell filling problem. In *Perspectives on morphological organization*, pages 139—-158. Brill.

Su Lin Blodgett, Lisa Green, and Brendan O'Connor. 2016. Demographic dialectal variation in social media: A case study of African-American English. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*.

Monojit Choudhury and Amit Deshpande. 2021. How linguistically fair are multilingual pre-trained language models? In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Bernard Comrie, Martin Haspelmath, and Balthasar Bickel. 2015. Leipzig glossing rules. *Conventions for Interlinear Morpheme-by-Morpheme Glosses. Leipzig: Max Planck Institute for Evolutionary Anthropology*.

Ewa Czaykowska-Higgins. 2009. Research models, community engagement, and linguistic fieldwork: Reflections on working within canadian indigenous communities. *Language documentation & conservation*, 3(1):182–215.

Dorottya Demszky, Devyani Sharma, J. Clark, Vinodkumar Prabhakaran, and Jacob Eisenstein. 2021. Learning to recognize dialect features. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Joel Dunham, Gina Cook, and Joshua Horner. 2014. LingSync & the online linguistic database: New models for the collection and management of data for language communities, linguists and language learners. In *Proceedings of the 2014 Workshop on the Use of Computational Methods in the Study of Endangered Languages*.

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.

Abteen Ebrahimi and Katharina Kann. 2021. How to adapt your pretrained multilingual model to 1600 languages. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*.

Kawin Ethayarajh and Dan Jurafsky. 2020. Utility is in the eye of the user: A critique of NLP leaderboards. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*.

Fahim Faisal, Sharlina Keshava, Md Mahfuz Ibn Alam, and Antonios Anastasopoulos. 2021. Sd-qa: Spoken dialectal question answering for the real world. In *Findings of the Association for Computational Linguistics: EMNLP 2021*.

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.

Clarissa Forbes, Garrett Nicolai, and Miikka Silfverberg. 2021. An FST morphological analyzer for the Gitksan language. In *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*.

Ryan Georgi. 2014. *From Aari to Zulu : massively multilingual creation of language tools using interlinear glossed text*. Ph.D. thesis, University of Washington.

Martin Haspelmath, Matthew S Dryer, David Gil, and Bernard Comrie. 2005. *The world atlas of language structures*. Oxford University Press.

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

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Sachin Kumar, Antonios Anastasopoulos, Shuly Wintner, and Yulia Tsvetkov. 2021. Machine translation into low-resource language varieties. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*.

William Lane and Steven Bird. 2020. Bootstrapping techniques for polysynthetic morphological analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

William D Lewis and Fei Xia. 2010. Developing odin: A multilingual repository of annotated language data for hundreds of the world's languages. *Literary and Linguistic Computing*, 25(3):303–319.

Patrick Littell, Anna Kazantseva, Roland Kuhn, Aidan Pine, Antti Arppe, Christopher Cox, and Marie-Odile Junker. 2018. Indigenous language technologies in Canada: Assessment, challenges, and successes. In *Proceedings of the 27th International Conference on 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.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Manuel Mager, Ximena Gutierrez-Vasques, Gerardo Sierra, and Ivan Meza-Ruiz. 2018. Challenges of language technologies for the indigenous languages of the Americas. In *Proceedings of the 27th International Conference on Computational Linguistics*.

Manuel Mager, Arturo Oncevay, Annette Rios, Ivan Vladimir Meza Ruiz, Alexis Palmer, Graham Neubig, and Katharina Kann, editors. 2021. *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*.

Arya D McCarthy, Christo Kirov, Matteo Grella, Amrit Nidhi, Patrick Xia, Kyle Gorman, Ekaterina Vylomova, Sabrina J Mielke, Garrett Nicolai, Miikka Silfverberg, et al. 2020a. Unimorph 3.0: Universal morphology. In *Proceedings of the 12th Language Resources and Evaluation Conference*.

Arya D. McCarthy, Rachel Wicks, Dylan Lewis, Aaron Mueller, Winston Wu, Oliver Adams, Garrett Nicolai, Matt Post, and David Yarowsky. 2020b. The Johns Hopkins University Bible corpus: 1600+ tongues for typological exploration. In *Proceedings of the 12th Language Resources and Evaluation Conference*.

Angelina McMillan-Major. 2020. Automating gloss generation in interlinear glossed text. *Proceedings of the Society for Computation in Linguistics*, 3(1):338–349.

Marianne Mithun. 2001. *The languages of native North America*. Cambridge University Press.

Sarah Moeller and Mans Hulden. 2018. Automatic glossing in a low-resource setting for language documentation. In *Proceedings of the Workshop on Computational Modeling of Polysynthetic Languages*.

Sarah Moeller, Ling Liu, Changbing Yang, Katharina Kann, and Mans Hulden. 2020. Igt2p: From interlinear glossed texts to paradigms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 5251–5262.

Garrett Nicolai and David Yarowsky. 2019. Learning morphosyntactic analyzers from the Bible via iterative annotation projection across 26 languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

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.

Alexis Palmer, Taesun Moon, and Jason Baldridge. 2009. Evaluating automation strategies in language documentation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2009 Workshop on Active Learning for Natural Language Processing*.

Bruce Rigsby. 1986. *Gitxsan Grammar*. University of Queensland.

Shruti Rijhwani, Antonios Anastasopoulos, and Graham Neubig. 2020. Ocr post-correction for endangered language texts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*.

Miikka Silfverberg and Mans Hulden. 2018. An encoder-decoder approach to the paradigm cell filling problem. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Till Speicher, Hoda Heidari, Nina Grgic-Hlaca, Krishna P Gummadi, Adish Singla, Adrian Weller, and Muhammad Bilal Zafar. 2018. A unified approach to quantifying algorithmic unfairness: Measuring individual &group unfairness via inequality indices. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*.

John Sylak-Glassman. 2016. The composition and use of the universal morphological feature schema (unimorph schema). *Johns Hopkins University*.

Daan van Esch, Ben Foley, and Nay San. 2019. Future directions in technological support for language documentation. In *Proceedings of the Workshop on Computational Methods for Endangered Languages*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Ekaterina Vylomova, Elizabeth Salesky, Sabrina Mielke, Gabriella Lapesa, Ritesh Kumar, Harald Hammarström, Ivan Vulić, Anna Korhonen, Roi Reichart, Edoardo Maria Ponti, and Ryan Cotterell, editors. 2021. *Proceedings of the Third Workshop on Computational Typology and Multilingual NLP*.

Monica Ward and Josef Genabith. 2003. Call for endangered languages: Challenges and rewards. *Computer Assisted language learning*, 16(2-3):233–258.

David Allen Wax. 2014. Automated grammar engineering for verbal morphology. Master's thesis, University of Washington.

Martina Wiltschko. 2008. The syntax of non-inflectional plural marking. *Natural Language and Linguistic Theory*, 26(3):639–694.

Winston Wu, Garrett Nicolai, and David Yarowsky. 2020. Multilingual dictionary based construction of core vocabulary. In *Proceedings of the 12th Language Resources and Evaluation Conference*.

Olga Zamaraeva. 2016. Inferring morphotactics from interlinear glossed text: Combining clustering and precision grammars. In *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*.

Xingyuan Zhao, Satoru Ozaki, Antonios Anastasopoulos, Graham Neubig, and Lori Levin. 2020. Automatic interlinear glossing for under-resourced languages leveraging translations. In *Proceedings of the 28th International Conference on Computational Linguistics*.

A Sample IGT data

The first four lines of a sample text from the Gitksan interlinear glossed text corpus. This example is revised from initial publication in Forbes et al. (2017).

```
Dim   mehldi'y         wila wilhl    win  hii       hagun
dim   mehl-T-i-'y      wila wil=hl   win  hii       hogun
PROSP tell-T-TR-1SG.II MANR be/do=CN COMP initially toward
bekwhl              mismaaxwsxum            get    go'ohl
bekw=hl             CVC~maaxws-xw-m         get    go'o=hl
arrive.PL[-3.II]=CN PL~fallen.snow-VAL-ATTR people LOC[-3.II]=CN
ts'ebim        Gitwinhlguu'l gik'uuhl.
ts'ep-m        Gitwinhlguu'l gi-k'uuhl
community-ATTR Gitwinhlguu'l prior-year
```

I will tell about when the white men first came to Kitwancool long ago.

```
Ha'on   dii 'nekw hlidaa   bekwhl              get    dipun,
ha'on   dii 'nekw hli=da   bekw=hl             get    dip=un
not.yet FOC long  PART=SPT arrive.PL[-3.II]=CN people ASSOC=DEM.PROX
ii   sagaytgoodindiithl               hli  gedihl
ii   sagayt-gooda-in-diit=hl          hli  get-T=hl
CCNJ together-all.gone-CAUS2-3PL.II=CN PART people-T[-3.II]=CN
Gitwinhlguu'l.
Gitwinhlguu'l
Gitwinhlguu'l
```

Not long after these people arrived, they gathered together the people of Kitwancool.

```
Hasakdiit     dimt      mehldiit      win  hlaa  dim   sii
hasak-diit    dim=t     mehl-T-diit   win  hlaa  dim   sii
desire-3PL.II PROSP=3.I tell-T-3PL.II COMP INCEP PROSP new
ha'niijokt        go'ohl        win  t'aahl
ha-'nii-jok-t     go'o=hl       win  t'aa=hl
INS-on-dwell-3.II LOC[-3.II]=CN COMP sit[-3.II]=CN
galts'ephl                  Gitwinhlguu'l.
gal-ts'ep=hl                Gitwinhlguu'l
container-community[-3.II]=CN Gitwinhlguu'l
```

They wanted to tell about the new place where the village of Kitwancool is to be.

```
'Nit  sagootxwhl               "government" siwatdiit,
'nit  si-goot-xw=hl            *government  si-wa-T-diit
3.III CAUS1-heart-VAL[-3.II]=CN *government  CAUS1-name-T[-TR]-3PL.II
ii   dim   'nii wenhl           dim   jokhl
ii   dim   'nii wen=hl          dim   jok=hl
CCNJ PROSP on   sit.PL[-3.II]=CN PROSP dwell[-3.II]=CN
aluugiget         go'ohl        lax "reserve"
aluu-CV~get       go'o=hl       lax *reserve
clearly-PL~people LOC[-3.II]=CN on  *reserve
siwatdiit.
si-wa-T-diit
CAUS1-name-T[-TR]-3PL.II
```

The plan of the so-called government was that they will have Indian people live on a so-called reserve.

B 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 _ _ _ _
```

C Fairseq parameters

Model We use the Fairseq (Ott et al., 2019) model implementation of Transformer (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 512. We train with the Adam optimizer starting of the learning rate at 0.001. We chose the batch size (400) and maximum updates (20000) based on the highest accuracy on the development data.