GlossLM: A Massively Multilingual Corpus and Pretrained Model for Interlinear Glossed Text

Michael Ginn*[1] **Lindia Tjuatja***[2] **Taiqi He**[2] **Enora Rice**[1]
Graham Neubig[2] **Alexis Palmer**[1] **Lori Levin**[2]
[1]University of Colorado Boulder [2]Carnegie Mellon University
michael.ginn@colorado.edu lindiat@andrew.cmu.edu
∗ Equal contribution

Abstract

Language documentation projects often involve the creation of annotated text in a format such as **interlinear glossed text (IGT)**, which captures fine-grained morphosyntactic analyses in a morpheme-by-morpheme format. However, there are few existing resources providing large amounts of standardized, easily accessible IGT data, limiting their applicability to linguistic research, and making it difficult to use such data in NLP modeling.

We compile the largest existing corpus of IGT data from a variety of sources, covering over 450k examples across 1.8k languages, to enable research on crosslingual transfer and IGT generation. We normalize much of our data to follow a standard set of labels across languages.

Furthermore, we explore the task of automatically generating IGT in order to aid documentation projects. As many languages lack sufficient monolingual data, we pretrain a large multilingual model on our corpus. We demonstrate the utility of this model by finetuning it on monolingual corpora, outperforming SOTA models by up to 6.6%. Our pretrained model and dataset are available on Hugging Face.[1]

1 Introduction

With nearly half of the world's 7,000 languages considered endangered, communities of minoritized language speakers are working to preserve and revitalize their languages (Seifart et al., 2018). These efforts often involve collection, analysis, and annotation of linguistic data. Annotated text can be used in the creation of reference materials (such as dictionaries and grammars) as well as to develop language technologies including searchable digital text (Blokland et al., 2019; Rijhwani et al., 2023) and computer-assisted educational tools (Uibo et al., 2017; Chaudhary et al., 2023).

[1]https://huggingface.co/collections/lecslab/glosslm-66da150854209e910113dd87

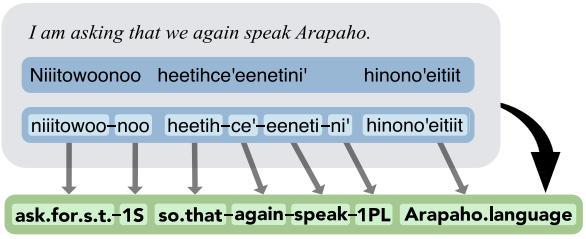

Figure 1: Components of interlinear gloss with an Arapaho sentence and English translation (Cowell, 2020). Blue boxes show transcriptions that are *unsegmented* (top) or *segmented* (bottom). Segmented text is split into morphemes which are aligned with the gloss labels shown in the green box. The task of automatic glossing uses some or all of the information in the gray box (transcription & translation) to generate the gloss line.

Interlinear glossed text (IGT) is a widespread format in language documentation for linguistic annotation. IGT is a multi-line data format (see Figure 1) which includes (1) a transcription of speech in the language, (2) an aligned morpheme-by-morpheme description, and oftentimes (3) a free translation. IGT can be used to illustrate morphosyntactic features of languages that other researchers may not be familiar with, and it is a popular format for examples in linguistics papers and textbooks. It also serves as a resource in the NLP context for the creation of morphological paradigms (Moeller et al., 2020), machine translation (Zhou et al., 2019), generating precision grammars (Bender et al., 2013), and other tools including POS taggers and dependency parsers (Georgi, 2016).

Compiling IGT Data Though IGT often follows a common glossing format, gloss conventions vary wildly. Furthermore, IGT data is rarely compiled into large, standardized corpora, often existing as scattered examples in research papers. To address this, we compile the largest corpus of digitized IGT from various existing sources, with over 450k examples in 1.8k languages (§3). We explore meth-

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 12267–12286
November 12-16, 2024

ods to normalize this data (§4)., standardizing over 80% of the grammatical glosses in the corpus to follow the UniMorph schema (Sylak-Glassman, 2016). We are releasing our corpora for future NLP, linguistics, and language documentation research.

Automating IGT Generation The creation of new IGT corpora is often difficult and time-consuming, requiring documenters to perform linguistic analysis and extensive documentation simultaneously. Research has found that computational tools can help accelerate annotation and overcome this bottleneck (Palmer et al., 2009) by predicting the gloss line of IGT given a transcription.

The majority of prior work on automatic glossing focuses on training monolingual models that can predict IGT for a single language (Moeller and Hulden, 2018; McMillan-Major, 2020; Zhao et al., 2020), however, these models can struggle when data is limited and require dedicated effort to train and deploy.

We aim to overcome the monolingual data bottleneck by creating a **multilingual pretrained glossing model** that can be adapted to specific languages and gloss formats. We continually pretrain a model on our corpus, and find that the pretrained multilingual model retains high performance across languages. We then finetune the pretrained model on monolingual data, including languages that do not appear in the pretraining corpus. Our models achieve new SOTA performance on five out of seven languages, and demonstrate clear improvements for low-resource language settings over an equivalent finetuned model without our continual pretraining (§7).

2 Interlinear Glossed Text (IGT)

2.1 Format

Interlinear glossed text is a structured data format which presents text in a language being studied along with *morphological glosses*—aligned labels that indicate each morpheme's meaning and/or grammatical function. Often, a free translation in a widely-spoken language is included as well. An IGT example for Arapaho is given in item 1 (Cowell, 2020), with glosses and translations in English.

(1) nuhu' tih-'eeneti-3i' heneenei3oobei-3i'
this when.PAST-speak-3PL IC.tell.the.truth-3PL
"When they speak, they tell the truth."

This example is *segmented*, with morphemes separated by dashes. Each morpheme in the Arapaho sentence (e.g. *tih*) is directly aligned with a gloss (e.g. when.PAST) that describes the morpheme's function and/or meaning. Labels in all caps (e.g. PAST) are grammatical glosses; lowercase labels (e.g. speak) are lexical glosses. Periods are used for *fusional morphemes*, which carry several morphological or lexical functions in a single morpheme.

IGT examples may instead use *unsegmented* transcriptions, as in the Uspanteko example in item 2 (Pixabaj et al., 2007).

(2) o sey xtok rixoqiil
o sea COM-buscar E3S-esposa
"O sea busca esposa."

Here, words and their labels are aligned, but no explicit alignment between morpheme glosses and individual morphemes is provided, and thus the segmentation of words into morphemes is unclear.

2.2 Challenges with Interlinear Glossing

Effective glossing requires expert knowledge of the target language and linguistic understanding of morphological patterns. Furthermore, certain factors exist that make this task particularly difficult for automated systems. Often, transcriptions are not segmented into morphemes, and systems must perform simultaneous segmentation and glossing.

Glossing conventions and formats may vary widely from documenter to documenter (Chelliah et al., 2021), with differences in label spelling (e.g. SING/SG/S to denote singular), formatting and punctuation, and language-specific labels. Furthermore, nearly all languages have very little digitized IGT data, posing difficulty to automated systems and human annotators alike. Finally, even when automated systems have been created, practical deployment remains an additional challenge for documenters.

3 GlossLM Corpus

While various publicly available sources of digitized IGT exist, the lack of unified data formatting and ease of access is a roadblock to using these resources effectively. To solve this problem, we compile and clean the largest IGT dataset from a variety of sources and languages. In total, our dataset contains over 450k IGT instances (from 250k unique sentences) in 1.8k languages, collected from six different IGT corpora. All sources are publicly available under the CC BY 4.0 License, allowing free use and redistribution, and we have confirmed with the creators of each source that our usage is

within the intended use. We make our artifacts available under the same license.

3.1 Data Sources

Corpus	Languages	IGT instances
ODIN	936	83,661
SIGMORPHON	7	68,604
IMTVault	1,116	79,522
APICS	76	15,805
UraTyp	35	1,719
Guarani Corpus	1	803
Total	1,782	250,582

Table 1: Number of unique examples and languages in each source corpus for the GLOSSLM dataset.

ODIN The Online Dictionary of Interlinear Text (ODIN, Lewis and Xia 2010) is a large dataset of 158k IGT examples representing 1496 languages, compiled by scraping IGT from linguistics documents on the internet. We use the preprocessed version of ODIN by He et al. (2023), which discards languages with fewer than five IGT samples, resulting in 84k unique glossed sentences across 936 languages.

SIGMORPHON Shared Task We use the IGT data from the 2023 SIGMORPHON Shared Task on Interlinear Glossing (Ginn et al., 2023). The data covers seven languages with diverse features and includes 69k glossed sentences. We use the shared task corpora as our primary evaluation sets, with the same splits as the shared task.

IMTVault IMTVault (Nordhoff and Krämer, 2022) is a recent aggregation of IGT data extracted from LaTeX code in books published by the Language Science Press. We use the 1.1 release (Nordhoff and Forkel, 2023) which includes 1116 languages and 80k examples.

APiCS The Atlas of Pidgin and Creole Language Structures (APiCS) is a set of books detailing grammatical features of 76 pidgin and creole languages (Michaelis et al., 2013a,b). APiCS online provides interactive versions of the books, including 16k IGT examples.

UraTyp UraTyp (Norvik et al., 2022) provides grammatical and typological information, collected from linguistic questionnaires on various languages. This includes a small number of IGT examples (1.7k) spanning 35 languages.

Guarani Corpus The Guarani Corpus (Maria Luisa Zubizarreta, 2023) consists of 803 examples of IGT, representing fifteen stories, for Guarani, a Tupian language spoken in South America. We use Beautiful Soup[2] to parse examples from HTML.

3.2 Preprocessing

In total we have 250k unique IGT instances in 1.8k languages. If datasets explicitly indicate whether an IGT line is segmented, we use this value. Otherwise, we determine segmentation by checking if a line has any morpheme boundary markers (the hyphen "-"). For segmented words, we remove the segmentation markers to create an additional unsegmented version of the same example, for a total of 451k examples (206k segmented).

We run `langid` (Lui and Baldwin, 2012) on translations to verify the translation language labels, and leave the translation field blank if the predicted language did not match the language indicated by the original source. Finally, we pad any non-lexical punctuation with spaces and normalize spacing, as our experiments indicate that our models are sensitive to this formatting.

3.3 Language Coverage

Within our dataset, around 90% of examples have an associated Glottocode (Hammarström et al., 2023), amounting to 1,785 unique Glottocodes and over 150 language families represented. While it would be ideal to have a relatively balanced set across languages and language families, many languages only have a few lines of IGT available, and thus our dataset has a long tail distribution across languages. The language with the greatest representation by far is Arapaho (from the SIGMORPHON Shared Task dataset) with almost 98k IGT instances, making up about 20% of the entire dataset. Overall, 25% of languages have fewer than 5 IGT instances, 50% have fewer than 10, and 75% have fewer than 54. We include histograms for the distributions across languages and language families in Appendix A, as well as preliminary analysis of typological coverage using the Grambank database (Skirgård et al., 2023) in Appendix B.

4 Normalizing Gloss Labels

4.1 Motivation

As our data comes from a variety of sources, spanning many languages and documentation projects,

[2] https://pypi.org/project/beautifulsoup4/

there is a great amount of diversity in the morphological glosses used. This includes cases where several different labels are used to indicate the same feature (e.g. SING, SG, or S for singular), as well as formatting differences such as the usage of periods (e.g. 1SG vs 1.SG).

We explore the feasibility and value of normalizing glosses to a single standardized format. On one hand, normalizing glosses may make it easier to train models that utilize crosslingual information through shared gloss labels, but on the other hand, it is difficult (perhaps impossible) to select a single schema that preserves the original intent of all annotators.

We split gloss lines by period and count the number and frequency of unique grammatical gloss labels across our corpus (focusing on the all-caps functional glosses, not stem translations) and visualize the distribution in Figure 2.

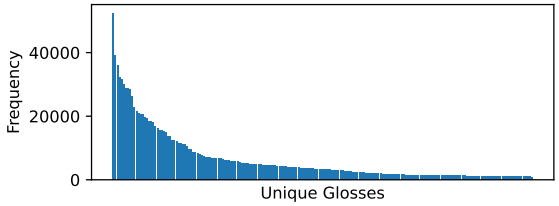

Figure 2: Distribution of unique glosses across all languages.

There are 11,493 unique glosses which roughly form a Zipfian distribution (Zipf, 1945). The most common glosses, unsurprisingly, are labels such as PL (plural, 52,488 instances), 3SG (3rd-person singular, 39,147), and PAST (36,124), which occur broadly across many languages.

Normalizing all unique glosses would be a monumental task with uncertain benefits. However, we observe that the 200 most common gloss types account for 82.7% of glosses in our dataset. We focus on normalizing these glosses: e.g. all instances of PAST and PST, which are both in the top 200, should be normalized to the same label. We note that there are other aspects of the data that vary (e.g. periods vs. underscores for multi-word glosses, representing non-concatenative morphology), as addressed in Mortensen et al. (2023). Future work could potentially focus on the benefits of these aspects of normalization on training.

4.2 Methodology

We select the UniMorph schema of Sylak-Glassman (2016) as our standardized set. While no single set of labels captures the intricacies of all of the world's languages, UniMorph is widely used and has coverage for many common features.

The two lead authors of this paper jointly created a mapping from the labels in our dataset to UniMorph labels. While many mappings were obvious (or already compatible with UniMorph), others posed a myriad of issues. For glosses primarily used for a single language, we consulted the original source dataset to determine the meaning.

Ambiguous labels Several labels were ambiguous, corresponding to one of several UniMorph glosses depending on the language and annotator. For example, the label S (appearing 20,855 times) is used for singular, subject, and noun/sustantivo (at least) in our dataset. In order to map these, we would need to analyze their meaning on a sentence-by-sentence basis, which was not feasible; thus, we left such ambiguous labels as-is.

Glosses not in UniMorph UniMorph primarily focuses on common crosslingual inflectional features, and does not cover the full extent of the morphological systems of the world. We observed 64 of 200 (32%) gloss labels with no clear UniMorph equivalent, including demonstratives (DEM, 15,585 instances), obliques (OBL, 13,639), and clitics (CL, 7,453). In many cases, there is a related UniMorph gloss that is more general or more specific; for example preterite (PRET, 1,986) could be mapped to simple past (PST). However, this would be an imprecise mapping, and could be confusing to a linguist of the particular language, so we again elect to leave these glosses unmapped.

We use our mapping to normalize the dataset and make the normalized version available in addition to the original.

4.3 Use in Future Research

We believe our dataset can potentially be useful across NLP research, linguistic research, and language documentation. NLP researchers benefit from a single, easily-accessible dataset covering many languages, which can be used for future research on interlinear glossing, morpheme analysis, segmentation, and translation.

Linguistics researchers will be able to use the dataset to easily search for phenomena across lan-

guages, particularly with the normalized version of the dataset. For example, a linguist could find examples of sentences demonstrating the ergative/absolutive distinction. They could further refine this analysis by narrowing the results to a set of related languages, using the glottocodes in our dataset.

As another example, if a linguist wished to determine how prior researchers have annotated examples in a particular language, they would previously have to search across research papers, textbooks, and small corpora. With our dataset, it is trivial to pull up all of the examples in a particular language, potentially compiled from many sources.

5 Automatic IGT Generation

Next, we evaluate the applicability of our dataset to the NLP task of automatic gloss prediction. We select the IGT data from the SIGMORPHON Shared Task on Interlinear Glossing (Ginn et al., 2023) to use for evaluation and testing, as this data consistently adheres to a set of glossing conventions and has been evaluated on prior models.

5.1 Target Languages

We reuse the train/eval/test splits for the seven languages from the SIGMORPHON Shared Task (Table 2). We designate three languagues—Arapaho, Tsez, and Uspanteko—as *in-domain languages*, which are included in the GLOSSLM corpus. We use Gitksan, Lezgi, Natugu, and Nyangbo as *out-of-domain languages*, which are omitted from the corpus. All languages except Nyangbo include translations.

The shared task included two distinct settings:

- In the *open track*, the transcription lines were segmented into morphemes. This becomes a token classification task, and tends to be far easier, with SOTA models achieving 80-90% accuracy (and even a naïve method that simply selects the most common gloss for each morpheme was very effective).
- In the *closed track*, transcription lines were not segmented. This setting is far more challenging, as models must jointly learn to segment words and predict glosses, and the best models achieved as low as 11% accuracy on small datasets. On the other hand, these models have the potential to be more valuable to documentation projects, where segmented text may not be available.

In our experiments, we focus on the unsegmented setting (closed track). However, the segmented data is also included in the GLOSSLM corpus, and could easily be used in future research.

		Finetuning		
Language	Pre-train	Train	Eval	Test
Other languages	198,121	-	-	-
In-domain languages				
Arapaho (arp)	39,132	39,132	4,892	4,892
Tsez (ddo)	3,558	3,558	445	445
Uspanteko (usp)	9,774	9,774	232	633
Out-of-domain languages				
Gitksan (git)	-	74	42	37
Lezgi (lez)	-	705	88	87
Natugu (ntu)	-	791	99	99
Nyangbo (nyb)	-	2,100	263	263

Table 2: Number of total (unsegmented) pretrain (for in-domain languages), train, evaluation, and test examples for the target languages.

5.2 Evaluation Metrics

For evaluating predictions, we strip punctuation (except for within glosses).[3] We evaluate **morpheme accuracy**, which counts the number of correct morpheme glosses *in the correct position*. Hence, if a gloss is incorrectly inserted or deleted, the subsequent glosses will be incorrect. We also evaluate **word accuracy**, which counts the number of entire correct word glosses.

However, accuracy may sometimes be too strict of a measurement—especially for generative models—as minor character insertions/deletions in the label are penalized heavily. Thus, we also evaluate **chrF++** (Popović, 2015), a character-level metric often used in machine translation. chrF++ measures the F1 score over character n-grams between the reference and predictions, and is robust to insertions and deletions, unlike accuracy.

6 GlossLM Model

Using our IGT corpus described in section 3, we train a single multilingual pretrained model for the glossing task that can be easily adapted to documentation projects, for both seen languages and unseen ones.

6.1 Architecture

We use the ByT5 model, a multilingual pretrained model using the T5 architecture (Raffel et al.,

[3] Because of this post-processing, our results for baselines are slightly different than what original sources report.

2020). ByT5 operates on byte-level inputs, as opposed to word or subword tokens, making it easily adaptable to different scripts and unseen languages. We use the ByT5-base model (582M parameters), pretrained on the mC4 dataset (Xue et al., 2021). We did not experiment with pretraining a randomly initialized model, as pretraining runs are expensive and we predict that the pretrained non-IGT base model serves as a better initialization.

6.2 IGT Pretraining

We continually pretrain the ByT5 model on the GlossLM corpus described in section 3. As we are evaluating on unsegmented IGT data, we omit segmented data for the evaluation languages from the pretraining corpus.[4] We structure the glossing task as a text-to-text problem, training the model with examples formatted with the following prompt:

```
Provide the glosses for the following
transcription in <lang>.

Transcription in <lang>: <transcription>
Transcription segmented: <yes/no/unknown>
Translation in <metalang>: <translation>
Glosses:
```

Models are trained to output the gloss line following the above prompt input. We include translations, which has been shown to provide benefits in gloss prediction (Zhao et al., 2020; Ginn et al., 2023). For some data, a translation was not available ($\approx 3\%$ of the training data), in which case the translation line is omitted. We pretrain models using the hyperparameters given in Appendix C. We did not conduct hyperparameter search, only tuning the batch size, to fit in our GPUs, and epochs and early stopping, to ensure convergence.

6.3 Performance of Pretrained Model

When training massively multilingual models, performance on individual languages can sometimes degrade in what is dubbed the "curse of multilinguality" (Conneau et al., 2020; Chang et al., 2023). To investigate this issue, we evaluate our pretrained model on the in-domain languages without any additional finetuning.

We compare the performance of our pretrained model to the current SOTA, which is the second system from Girrbach (2023a), as shown in Figure 3. We find that the model outperforms the SOTA across all three in-domain languages. This result gives little evidence to believe our model suffers from the curse of multilinguality, as it retains good performance across several languages.[5]

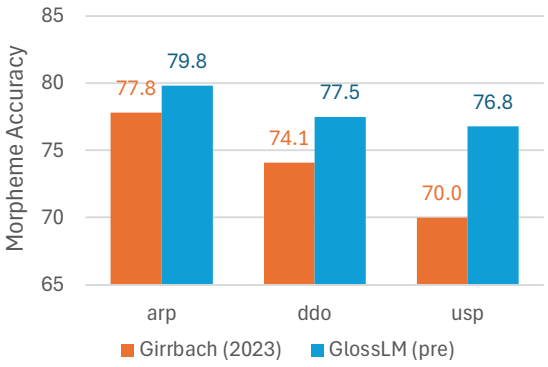

Figure 3: Comparison of our pretrained model and the SOTA (Girrbach, 2023a) for in-domain languages on unsegmented data. Our model outperforms on all three languages.

Our pretrained model can be used for automated glossing across several languages, without needing to train and serve separate monolingual models. This could be valuable to real-world documentation projects, as we can serve a single pretrained model that can be used across projects, significantly reducing the barrier to using an automated system.

The languages evaluated here are well-represented in the pretraining corpus, from Tsez (3.7k unsegmented examples, about 3% of the total corpus) to Arapaho (39.1k examples, 21%). A natural question is whether the model retains good performance on a language which occurs very rarely in the pretraining corpus. We simulate this scenario by adding a small amount of data to the pretraining corpus for two unseen languages: 1000 examples in Nyangbo and 500 in Natugu (less than 1% of the total corpus). We evaluate on the unseen test split and observe 76.8% and 55.0% morpheme accuracy, respectively. These results indicate the model can still perform well on languages that are sparse in the pretraining corpus.

---

[4] However, our corpus includes segmented IGT examples in other languages, which we do not evaluate.

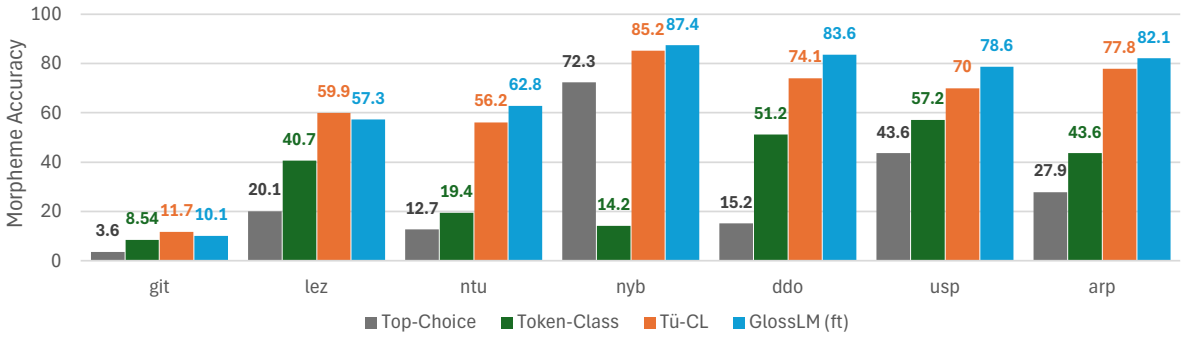

Figure 4: Morpheme accuracy for various systems.

7 Results

7.1 Comparison with Baselines

After pretraining GLOSSLM, we train finetuned versions for each of the languages in the test set. We first describe our finetuning procedure and compare results of our finetuned models against baselines from Ginn (2023) (§7.1). Then, to further isolate the efficacy of pretraining, we compare the finetuned versions of GLOSSLM to a finetuned ByT5 model without the multilingual gloss pretraining (§7.2). Finally, we explore whether training on a minimally normalized version of the data improves performance (§7.3).

As previously noted, we focus on the *unsegmented* setting; for completeness, we provide full results for both segmented and unsegmented data in Appendix D.

Finetuning can help align the model to a particular language or even a new unseen language. We finetune our GLOSSLM pretrained model on the training dataset for each language individually, and label these runs as GLOSSLM$_{\text{FT}}$. We do this for both the in-domain languages, to focus the model on a single language, as well as the out-of-domain languages, allowing us to study the model's adaptation to unseen languages.

Finetuning used the same parameters as pretraining, but with 100 training epochs and early stopping (patience 3, start epoch 15)[6], and took anywhere from 20 minutes to one day for each language. Inference uses beam search with $n = 3$ beams.

We compare the finetuned GLOSSLM models with three baseline systems which include the state-of-the-art from prior work:

1. **TOP-CHOICE** selects the most frequent label associated with each morpheme/word in the training data, and assigns "???" to unseen morphemes.

2. **TOKEN-CLASS** treats the glossing task as a token classification problem, where the output vocabulary consists of the IGT morpheme or word-level labels. Each target language's data is used to train a language-specifc TOKEN-CLASS model, which uses the RoBERTa architecture with default hyperparameters without any additional pretraining (Liu et al., 2019). This was used as the baseline model for the SIGMORPHON 2023 Shared Task on interlinear glossing (Ginn, 2023).

3. **TÜ-CL** (Girrbach, 2023b) uses straight-through gradient estimation (Bengio et al., 2013) to induce latent, discrete segmentations of input texts, and predicts glosses using a multilayer perceptron.

As shown in Figure 4, we find that our finetuned models outperform SOTA in all but two languages (Gitksan and Lezgi). The TÜ-CL model (Girrbach, 2023a), which is a close second and outperforms on Gitksan and Lezgi, uses explicit latent segmentations, which seems to be particularly beneficial for the very low-resource, unsegmented setting.

To illustrate common errors from the finetuned GLOSSLM moderls, we include examples of system outputs in Table 10. When inspecting outputs, we observe that there are sometimes inconsistencies in the IGT labels where multiple interchangeable glosses are used for the same morpheme. While we try to account for a portion of grammatical gloss variations (§4, §7.3), this is

[5]These languages do make up large fractions of our pretraining corpus, so the model will almost certainly underperform on underrepresented languages.

[6]For Gitksan, due to the size of the training set, we set max epochs to 300 and patience to 15.

particularly an issue for lexical glosses (e.g. the Arapaho word *'eeneisih'i* is glossed in the data as "how.X.things.are.named", "how.s.o..is.named", and "how.things.are.named"). We also find outputs that include lexical items present in the translation that are not included in the gold gloss, indicating that the model may rely too heavily on translations when predicting lexical glosses in certain cases.

7.2 Comparison with Finetuned ByT5

To directly assess the impact of pretraining on performance, we finetune ByT5 models on each language in the test set with the same configuration as for the finetuned GLOSSLM models. We then compare the performance of our models (which have undergone both multilingual gloss pretraining and finetuning) with analogous ByT5 models (without multilingual gloss pretraining), as shown in Figure 5.

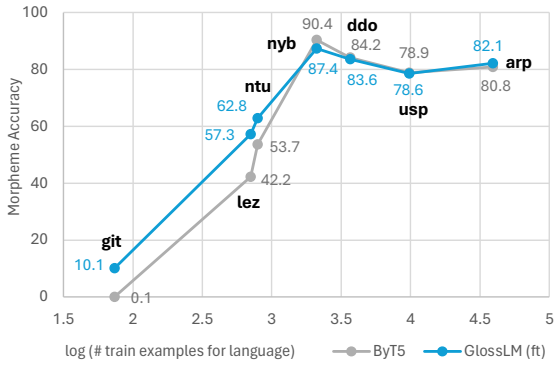

Figure 5: Performance after monolingual finetuning, comparing a standard pretrained ByT5 with a continually pretrained GLOSSLM model. The x-axis uses the log (base 10) of the number of training examples in a given language, for readability.

We observe mixed results, which are largely dependent on the size of the training corpus. For languages with less training data (Gitksan, Lezgi, Natugu) the GLOSSLM$_{FT}$ model outperforms the finetuned ByT5 model (by 10.0, 15.1, and 9.1 points respectively). In the case of Gitksan, the finetuned ByT5 model is completely unable to produce well-formatted output (likely due to the tiny training corpus) while the GLOSSLM$_{FT}$ model does not struggle with this as much. A possible explanation is that even if there are no similar languages in the pretraining corpus, the GLOSSLM$_{FT}$ can leverage knowledge about IGT formatting from unrelated languages.

For Lezgi—which shows the greatest improvements from pretraining with the GLOSSLM

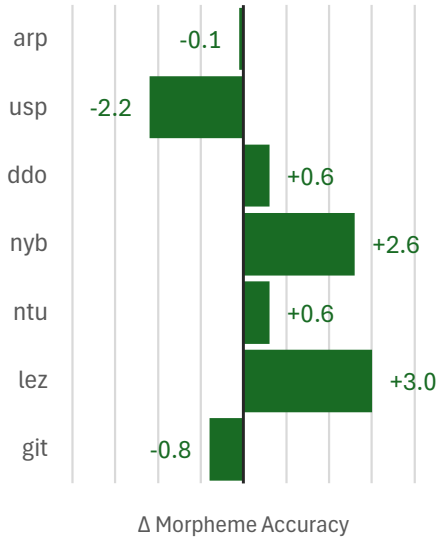

Figure 6: Change in morpheme accuracy after normalizing glosses to the UniMorph schema and finetuning GLOSSLM. We observe small improvements for several languages, but worse performance in two cases.

corpus—a qualitative analysis of examples with the greatest morpheme error rate between finetuned ByT5 and GLOSSLM reveals that there are regular error patterns that are fixed with continual pretraining. For example, the finetuned ByT5 model often outputs AOC instead of AOR and OLB instead of ERG, whereas the finetuned GLOSSLM gets these correct. We include examples of these outputs in Table 11.

However, with enough data (starting with Nyangbo, 2100 examples) the two approaches achieve nearly identical performance. This is an unsurprising result, indicating that large amounts of high-quality monolingual data overshadow any benefits from crosslingual transfer. Furthermore, we note that benefits of multilingual gloss pretraining shown may not be unique to the T5 architecture—while we only experiment with ByT5, our pretraining strategy could be applied to other architectures.

7.3 Effect of Gloss Normalization

Finally, we experiment with pretraining and finetuning on a minimally normalized version of the dataset, where the 200 most frequent grammatical labels are mapped to a set of standard labels.

We repeat the same pretraining and finetuning process as before. When comparing the performance of the pretrained model before finetuning on in-domain languages, we find minimal differences compared to the results in §6.3.

We report the change in morpheme accuracy af-

ter normalizing in Figure 6. We observe mixed results. Some languages (Arapaho, Uspanteko, Gitksan) show worse or equivalent performance. Others (Tsez, Nyangbo, Natugu, and Lezgi) show small to moderate improvements, with Lezgi achieving the largest improvement of 3.0 percentage points. Thus, we found that normalization was most helpful when finetuning the pretrained model on unseen languages with a low-to-moderate amount of training data.

8 Related Works

Automatic Interlinear Glossing Recent research has explored various methods for generating IGT, including rule-based methods (Bender et al., 2014), active learning (Palmer et al., 2010, 2009), conditional random fields (Moeller and Hulden, 2018; McMillan-Major, 2020), and neural models (Moeller and Hulden, 2018; Zhao et al., 2020). The 2023 SIGMORPHON Shared Task (Ginn et al., 2023) compared a number of highly-effective IGT generation systems, including ensembled LSTMs (Coates, 2023), straight-through gradient estimation (Girrbach, 2023a), CRF-neural systems (Okabe and Yvon, 2023a), and BiLSTM encoders (Cross et al., 2023).

In particular, this work is inspired by He et al. (2023), which pretrains ByT5 models on the ODIN corpus, and Okabe and Yvon (2023b), which pretrains a CRF model on the IMTVault corpus. However, neither explore using a pretraining corpus as large as ours, nor do they evaluate on unsegmented text. Furthermore, neither of these studies find significant benefits to using pretraining corpora, while we observe benefits under certain conditions.

Large Multilingual Pretrained Models Prior work has shown that large, massively multilingual pretrained language models can boost performance across low- and high-resource languages on a variety of tasks. These include encoder-decoder models trained with the masked language modeling objective (Pires et al., 2019; Conneau et al., 2020) and the span corruption objective (Xue et al., 2021, 2022), as well as decoder-only language models (Workshop et al., 2023; Shliazhko et al., 2024). Work such as Wang et al. (2020), Adelani et al. (2022b), and Adelani et al. (2022a) has shown that continual pretraining and/or finetuning large multilingual models is an effective method for tasks like low-resource language translation and named entity recognition.

9 Conclusion

High-quality language documentation involves an incredible amount of effort. We compile, normalize, and release the largest corpus of multilingual IGT data, enabling future research in linguistics, NLP, and documentation. Furthermore, we demonstrate the applicability of our corpus by pretraining a multilingual neural model for automatic generation of IGT. We finetune the model on monolingual corpora, showing benefits on low-resource languages due to multilingual pretraining. In five out of seven languages, we achieve a new SOTA on automatic IGT generation.

Limitations

Our work evaluates the effectiveness of massively multilingual pretraining on seven IGT datasets in different languages using the ByT5 architecture. We did not experiment with pretraining on other architectures, which may show different results. While we believe the selected evaluation languages cover a diverse set of features and dataset sizes, other languages may show better or worse results.

Our pretraining corpus consists of all IGT data we were able to find and utilize. As such, it is not evenly distributed among languages, overrepresenting a few languages with large language documentation efforts. Thus, models pretrained on the corpus will perform better on these and similar languages.

The only hyperparameter optimization we performed was finding a batch size that fit our GPUs and tuning epochs and early stopping in order to ensure convergence. We did not conduct hyperparameter search over other parameters such as learning rate or optimizer, architecture parameters, or dataset splits.

When evaluating predictions, we ignored punctuation (as our primary concern was gloss accuracy). Certain models may perform better or worse at outputting proper punctuation format, which could be a concern for certain applications.

Finally, it has been demonstrated that IGT generation models are often not robust to domain shift, compared with human annotators (Ginn and Palmer, 2023). Our models will likely have impacted performance for out-of-domain texts, such as highly technical or domain-specific language.

Ethics Statement

We hope this work can aid in the struggle against language extinction. However, language documentation, preservation, and revitalization require far more than generating IGT, and we should be careful not to understate the difficulty of these efforts. We utilize datasets produced by the painstaking effort of language documenters and speakers, and strive to treat the corpora as human artifacts, not just data to be consumed.

We hope our research can aid documentary linguists through automated gloss prediction. However, we caution against using these systems without human collaboration, as they can introduce error and miss novel linguistic insights. There is some risk of these systems being used to replace human annotators, which we strongly oppose.

While we try to train only the necessary models for our experiments, training large machine learning models carries an environmental cost (Bender et al., 2021; Strubell et al., 2020).

We do not evaluate our corpus for bias (racial, gender, etc) or inclusive language, and it's possible that our models can carry some of these biases.

Finally, NLP work that involves Indigenous and endangered languages has historically been plagued by colonialist approaches to data use and technology development (Schwartz, 2022). The large IGT datasets for endangered languages (Arapaho, Guarani, Uspanteko) were collected in collaboration with native communities, and our work is in accordance with the agreements for their usage.

Acknowledgements

We would like to thank the three anonymous reviewers and meta-reviewer for their helpful feedback on this work. Parts of this work were supported by the National Science Foundation under Grant No. 2149404, "CAREER: From One Language to Another" and Grant No. 2211951, "From Acoustic Signal to Morphosyntactic Analysis in One End-to-End Neural System." Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

David Adelani, Jesujoba Alabi, Angela Fan, Julia Kreutzer, Xiaoyu Shen, Machel Reid, Dana Ruiter, Dietrich Klakow, Peter Nabende, Ernie Chang, Tajuddeen Gwadabe, Freshia Sackey, Bonaventure F. P. Dossou, Chris Emezue, Colin Leong, Michael Beukman, Shamsuddeen Muhammad, Guyo Jarso, Oreen Yousuf, Andre Niyongabo Rubungo, Gilles Hacheme, Eric Peter Wairagala, Muhammad Umair Nasir, Benjamin Ajibade, Tunde Ajayi, Yvonne Gitau, Jade Abbott, Mohamed Ahmed, Millicent Ochieng, Anuoluwapo Aremu, Perez Ogayo, Jonathan Mukiibi, Fatoumata Ouoba Kabore, Godson Kalipe, Derguene Mbaye, Allahsera Auguste Tapo, Victoire Memdjokam Koagne, Edwin Munkoh-Buabeng, Valencia Wagner, Idris Abdulmumin, Ayodele Awokoya, Happy Buzaaba, Blessing Sibanda, Andiswa Bukula, and Sam Manthalu. 2022a. A few thousand translations go a long way! leveraging pre-trained models for African news translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3053–3070, Seattle, United States. Association for Computational Linguistics.

David Adelani, Graham Neubig, Sebastian Ruder, Shruti Rijhwani, Michael Beukman, Chester Palen-Michel, Constantine Lignos, Jesujoba Alabi, Shamsuddeen Muhammad, Peter Nabende, Cheikh M. Bamba Dione, Andiswa Bukula, Rooweither Mabuya, Bonaventure F. P. Dossou, Blessing Sibanda, Happy Buzaaba, Jonathan Mukiibi, Godson Kalipe, Derguene Mbaye, Amelia Taylor, Fatoumata Kabore, Chris Chinenye Emezue, Anuoluwapo Aremu, Perez Ogayo, Catherine Gitau, Edwin Munkoh-Buabeng, Victoire Memdjokam Koagne, Allahsera Auguste Tapo, Tebogo Macucwa, Vukosi Marivate, Mboning Tchiaze Elvis, Tajuddeen Gwadabe, Tosin Adewumi, Orevaoghene Ahia, Joyce Nakatumba-Nabende, Neo Lerato Mokono, Ignatius Ezeani, Chiamaka Chukwuneke, Mofetoluwa Oluwaseun Adeyemi, Gilles Quentin Hacheme, Idris Abdulmumin, Odunayo Ogundepo, Oreen Yousuf, Tatiana Moteu, and Dietrich Klakow. 2022b. MasakhaNER 2.0: Africa-centric transfer learning for named entity recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4488–4508, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Emily M Bender, Joshua Crowgey, Michael Wayne Goodman, and Fei Xia. 2014. Learning grammar specifications from igt: A case study of chintang. In *Proceedings of the 2014 Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 43–53.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and*

Transparency, pages 610–623, Virtual Event Canada. ACM.

Emily M. Bender, Michael Wayne Goodman, Joshua Crowgey, and Fei Xia. 2013. Towards creating precision grammars from interlinear glossed text: Inferring large-scale typological properties. In *Proceedings of the 7th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 74–83, Sofia, Bulgaria. Association for Computational Linguistics.

Yoshua Bengio, Nicholas Léonard, and Aaron Courville. 2013. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*.

Rogier Blokland, Niko Partanen, Michael Rießler, and Joshua Wilbur. 2019. Using computational approaches to integrate endangered language legacy data into documentation corpora: Past experiences and challenges ahead. In *Workshop on Computational Methods for Endangered Languages, Honolulu, Hawai'i, USA*, volume 2, pages 24–30.

Tyler A. Chang, Catherine Arnett, Zhuowen Tu, and Benjamin K. Bergen. 2023. When is multilinguality a curse? language modeling for 250 high- and low-resource languages.

Aditi Chaudhary, Arun Sampath, Ashwin Sheshadri, Antonios Anastasopoulos, and Graham Neubig. 2023. Teacher perception of automatically extracted grammar concepts for L2 language learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3776–3793, Singapore. Association for Computational Linguistics.

Shobhana Chelliah, Mary Burke, and Marty Heaton. 2021. Using interlinear gloss texts to improve language description. *Indian Linguistics*, 82.

Edith Coates. 2023. An ensembled encoder-decoder system for interlinear glossed text. In *Proceedings of the 20th SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 217–221, Toronto, Canada. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Andrew Cowell. 2020. The arapaho lexical and text database. Department of Linguistics, University of Colorado. Boulder, CO.

Ziggy Cross, Michelle Yun, Ananya Apparaju, Jata MacCabe, Garrett Nicolai, and Miikka Silfverberg. 2023. Glossy bytes: Neural glossing using subword encoding. In *Proceedings of the 20th SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 222–229, Toronto, Canada. Association for Computational Linguistics.

Ryan Alden Georgi. 2016. *From Aari to Zulu: massively multilingual creation of language tools using interlinear glossed text*. Ph.D. thesis.

Michael Ginn. 2023. Sigmorphon 2023 shared task of interlinear glossing: Baseline model.

Michael Ginn, Sarah Moeller, Alexis Palmer, Anna Stacey, Garrett Nicolai, Mans Hulden, and Miikka Silfverberg. 2023. Findings of the SIGMORPHON 2023 shared task on interlinear glossing. In *Proceedings of the 20th SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 186–201, Toronto, Canada. Association for Computational Linguistics.

Michael Ginn and Alexis Palmer. 2023. Robust generalization strategies for morpheme glossing in an endangered language documentation context. In *Proceedings of the 1st GenBench Workshop on (Benchmarking) Generalisation in NLP*, pages 89–98, Singapore. Association for Computational Linguistics.

Leander Girrbach. 2023a. SIGMORPHON 2022 shared task on grapheme-to-phoneme conversion submission description: Sequence labelling for G2P. In *Proceedings of the 20th SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 239–244, Toronto, Canada. Association for Computational Linguistics.

Leander Girrbach. 2023b. Tü-CL at SIGMORPHON 2023: Straight-through gradient estimation for hard attention. In *Proceedings of the 20th SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 171–185, Toronto, Canada. Association for Computational Linguistics.

Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2023. Glottolog 4.8.

Taiqi He, Lindia Tjuatja, Nathaniel Robinson, Shinji Watanabe, David R. Mortensen, Graham Neubig, and Lori Levin. 2023. SigMoreFun submission to the SIGMORPHON shared task on interlinear glossing. In *Proceedings of the 20th SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 209–216, Toronto, Canada. Association for Computational Linguistics.

W. D. Lewis and F. 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.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019.

RoBERTa: A Robustly Optimized BERT Pretraining Approach. ArXiv:1907.11692 [cs].

Marco Lui and Timothy Baldwin. 2012. langid.py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 System Demonstrations*, pages 25–30, Jeju Island, Korea. Association for Computational Linguistics.

Maria Luisa Zubizarreta. 2023. Guarani corpus. `https://guaranicorpus.usc.edu/index.html`. Accessed: 2024-02-11.

Angelina McMillan-Major. 2020. Automating Gloss Generation in Interlinear Glossed Text. *Proceedings of the Society for Computation in Linguistics*, 3(1):338–349. Publisher: University of Mass Amherst.

Susanne Maria Michaelis, Philippe Maurer, Martin Haspelmath, and Magnus Huber, editors. 2013a. *The Atlas of Pidgin and Creole Language Structures*. Oxford University Press, Oxford.

Susanne Maria Michaelis, Philippe Maurer, Martin Haspelmath, and Magnus Huber, editors. 2013b. *The Survey of Pidgin and Creole Languages*. Oxford University Press, Oxford. 3 volumes. Volume I: English-based and Dutch-based Languages; Volume II: Portuguese-based, Spanish-based, and French-based Languages. Volume III: Contact Languages Based on Languages From Africa, Australia, and the Americas.

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*, pages 84–93, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

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 (EMNLP)*, pages 5251–5262, Online. Association for Computational Linguistics.

David R. Mortensen, Ela Gulsen, Taiqi He, Nathaniel Robinson, Jonathan Amith, Lindia Tjuatja, and Lori Levin. 2023. Generalized glossing guidelines: An explicit, human- and machine-readable, item-and-process convention for morphological annotation. In *Proceedings of the 20th SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 58–67, Toronto, Canada. Association for Computational Linguistics.

Sebastian Nordhoff and Robert Forkel. 2023. IMTVault.

Sebastian Nordhoff and Thomas Krämer. 2022. IMTVault: Extracting and enriching low-resource language interlinear glossed text from grammatical descriptions and typological survey articles. In *Proceedings of the 8th Workshop on Linked Data in Linguistics within the 13th Language Resources and Evaluation Conference*, pages 17–25, Marseille, France. European Language Resources Association.

Miina Norvik, Yingqi Jing, Michael Dunn, Robert Forkel, Terhi Honkola, Gerson Klumpp, Richard Kowalik, Helle Metslang, Karl Pajusalu, Minerva Piha, Eva Saar, Sirkka Saarinen, and Outi Vesakoski. 2022. Uralic Typological database - UraTyp.

Shu Okabe and François Yvon. 2023a. LISN @ SIGMORPHON 2023 shared task on interlinear glossing. In *Proceedings of the 20th SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 202–208, Toronto, Canada. Association for Computational Linguistics.

Shu Okabe and François Yvon. 2023b. Towards multilingual interlinear morphological glossing. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5958–5971, Singapore. Association for Computational Linguistics.

Alexis Palmer, Taesun Moon, and Jason Baldridge. 2009. Evaluating automation strategies in language documentation. In *Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing*, pages 36–44.

Alexis Palmer, Taesun Moon, Jason Baldridge, Katrin Erk, Eric Campbell, and Telma Can. 2010. Computational strategies for reducing annotation effort in language documentation: A case study in creating interlinear texts for Uspanteko. *Linguistic Issues in Language Technology*, 3.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Telma Can Pixabaj, Miguel Angel Vicente Méndez, María Vicente Méndez, and Oswaldo Ajcot Damián. 2007. *Text Collections in Four Mayan Languages*. Archived in The Archive of the Indigenous Languages of Latin America.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Shruti Rijhwani, Daisy Rosenblum, Michayla King, Antonios Anastasopoulos, and Graham Neubig. 2023. User-centric evaluation of OCR systems for kwak'wala. In *Proceedings of the Sixth Workshop*

on the Use of Computational Methods in the Study of Endangered Languages, pages 19–29, Remote. Association for Computational Linguistics.

Lane Schwartz. 2022. Primum Non Nocere: Before working with Indigenous data, the ACL must confront ongoing colonialism. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 724–731, Dublin, Ireland. Association for Computational Linguistics.

Frank Seifart, Nicholas Evans, Harald Hammarström, and Stephen C. Levinson. 2018. Language documentation twenty-five years on. *Language*, 94(4):e324–e345.

Oleh Shliazhko, Alena Fenogenova, Maria Tikhonova, Anastasia Kozlova, Vladislav Mikhailov, and Tatiana Shavrina. 2024. mgpt: Few-shot learners go multilingual. *Transactions of the Association for Computational Linguistics*, 12:58–79.

Hedvig Skirgård, Hannah J. Haynie, Damián E. Blasi, Harald Hammarström, Jeremy Collins, Jay J. Latarche, Jakob Lesage, Tobias Weber, Alena Witzlack-Makarevich, Sam Passmore, Angela Chira, Luke Maurits, Russell Dinnage, Michael Dunn, Ger Reesink, Ruth Singer, Claire Bowern, Patience Epps, Jane Hill, Outi Vesakoski, Martine Robbeets, Noor Karolin Abbas, Daniel Auer, Nancy A. Bakker, Giulia Barbos, Robert D. Borges, Swintha Danielsen, Luise Dorenbusch, Ella Dorn, John Elliott, Giada Falcone, Jana Fischer, Yustinus Ghanggo Ate, Hannah Gibson, Hans-Philipp Göbel, Jemima A. Goodall, Victoria Gruner, Andrew Harvey, Rebekah Hayes, Leonard Heer, Roberto E. Herrera Miranda, Nataliia Hübler, Biu Huntington-Rainey, Jessica K. Ivani, Marilen Johns, Erika Just, Eri Kashima, Carolina Kipf, Janina V. Klingenberg, Nikita König, Aikaterina Koti, Richard G. A. Kowalik, Olga Krasnoukhova, Nora L. M. Lindvall, Mandy Lorenzen, Hannah Lutzenberger, Tânia R. A. Martins, Celia Mata German, Suzanne van der Meer, Jaime Montoya Samamé, Michael Müller, Saliha Muradoglu, Kelsey Neely, Johanna Nickel, Miina Norvik, Cheryl Akinyi Oluoch, Jesse Peacock, India O. C. Pearey, Naomi Peck, Stephanie Petit, Sören Pieper, Mariana Poblete, Daniel Prestipino, Linda Raabe, Amna Raja, Janis Reimringer, Sydney C. Rey, Julia Rizaew, Eloisa Ruppert, Kim K. Salmon, Jill Sammet, Rhiannon Schembri, Lars Schlabbach, Frederick W. P. Schmidt, Amalia Skilton, Wikaliler Daniel Smith, Hilário de Sousa, Kristin Sverredal, Daniel Valle, Javier Vera, Judith Voß, Tim Witte, Henry Wu, Stephanie Yam, Jingting Ye, Maisie Yong, Tessa Yuditha, Roberto Zariquiey, Robert Forkel, Nicholas Evans, Stephen C. Levinson, Martin Haspelmath, Simon J. Greenhill, Quentin D. Atkinson, and Russell D. Gray. 2023. Grambank reveals the importance of genealogical constraints on linguistic diversity and highlights the impact of language loss. *Science Advances*, 9(16):eadg6175.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2020. Energy and policy considerations for modern deep learning research. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 13693–13696.

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

Heli Uibo, Jack Rueter, and Sulev Iva. 2017. Building and using language resources and infrastructure to develop e-learning programs for a minority language. *Proceedings of the Joint 6th Workshop on NLP for Computer Assisted Language Learning and 2nd Workshop on NLP for Research on Language Acquisition at NoDaLiDa*, 134:61–67.

Zihan Wang, Karthikeyan K, Stephen Mayhew, and Dan Roth. 2020. Extending multilingual BERT to low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2649–2656, Online. Association for Computational Linguistics.

BigScience Workshop, :, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, Dragomir Radev, Eduardo González Ponferrada, Efrat Levkovizh, Ethan Kim, Eyal Bar Natan, Francesco De Toni, Gérard Dupont, Germán Kruszewski, Giada Pistilli, Hady Elsahar, Hamza Benyamina, Hieu Tran, Ian Yu, Idris Abdulmumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jörg Frohberg, Joseph Tobing, Joydeep Bhattacharjee, Khalid Almubarak, Kimbo Chen, Kyle Lo, Leandro Von Werra, Leon Weber, Long Phan, Loubna Ben allal, Ludovic Tanguy, Manan Dey, Manuel Romero Muñoz, Maraim Masoud, María Grandury, Mario Šaško, Max Huang, Maximin Coavoux, Mayank Singh, Mike Tian-Jian Jiang, Minh Chien Vu, Mohammad A. Jauhar, Mustafa Ghaleb, Nishant Subramani, Nora Kassner, Nurulaqilla Khamis, Olivier Nguyen, Omar Espejel, Ona de Gibert, Paulo Villegas, Peter Henderson, Pierre Colombo, Priscilla Amuok, Quentin Lhoest, Rheza Harliman, Rishi Bommasani, Roberto Luis López, Rui Ribeiro, Salomey Osei, Sampo Pyysalo, Sebastian Nagel, Shamik Bose, Shamsuddeen Hassan Muhammad, Shanya Sharma,

Shayne Longpre, Somaieh Nikpoor, Stanislav Silberberg, Suhas Pai, Sydney Zink, Tiago Timponi Torrent, Timo Schick, Tristan Thrush, Valentin Danchev, Vassilina Nikoulina, Veronika Laippala, Violette Lepercq, Vrinda Prabhu, Zaid Alyafeai, Zeerak Talat, Arun Raja, Benjamin Heinzerling, Chenglei Si, Davut Emre Taşar, Elizabeth Salesky, Sabrina J. Mielke, Wilson Y. Lee, Abheesht Sharma, Andrea Santilli, Antoine Chaffin, Arnaud Stiegler, Debajyoti Datta, Eliza Szczechla, Gunjan Chhablani, Han Wang, Harshit Pandey, Hendrik Strobelt, Jason Alan Fries, Jos Rozen, Leo Gao, Lintang Sutawika, M Saiful Bari, Maged S. Al-shaibani, Matteo Manica, Nihal Nayak, Ryan Teehan, Samuel Albanie, Sheng Shen, Srulik Ben-David, Stephen H. Bach, Taewoon Kim, Tali Bers, Thibault Fevry, Trishala Neeraj, Urmish Thakker, Vikas Raunak, Xiangru Tang, Zheng-Xin Yong, Zhiqing Sun, Shaked Brody, Yallow Uri, Hadar Tojarieh, Adam Roberts, Hyung Won Chung, Jaesung Tae, Jason Phang, Ofir Press, Conglong Li, Deepak Narayanan, Hatim Bourfoune, Jared Casper, Jeff Rasley, Max Ryabinin, Mayank Mishra, Minjia Zhang, Mohammad Shoeybi, Myriam Peyrounette, Nicolas Patry, Nouamane Tazi, Omar Sanseviero, Patrick von Platen, Pierre Cornette, Pierre François Lavallée, Rémi Lacroix, Samyam Rajbhandari, Sanchit Gandhi, Shaden Smith, Stéphane Requena, Suraj Patil, Tim Dettmers, Ahmed Baruwa, Amanpreet Singh, Anastasia Cheveleva, Anne-Laure Ligozat, Arjun Subramonian, Aurélie Névéol, Charles Lovering, Dan Garrette, Deepak Tunuguntla, Ehud Reiter, Ekaterina Taktasheva, Ekaterina Voloshina, Eli Bogdanov, Genta Indra Winata, Hailey Schoelkopf, Jan-Christoph Kalo, Jekaterina Novikova, Jessica Zosa Forde, Jordan Clive, Jungo Kasai, Ken Kawamura, Liam Hazan, Marine Carpuat, Miruna Clinciu, Najoung Kim, Newton Cheng, Oleg Serikov, Omer Antverg, Oskar van der Wal, Rui Zhang, Ruochen Zhang, Sebastian Gehrmann, Shachar Mirkin, Shani Pais, Tatiana Shavrina, Thomas Scialom, Tian Yun, Tomasz Limisiewicz, Verena Rieser, Vitaly Protasov, Vladislav Mikhailov, Yada Pruksachatkun, Yonatan Belinkov, Zachary Bamberger, Zdeněk Kasner, Alice Rueda, Amanda Pestana, Amir Feizpour, Ammar Khan, Amy Faranak, Ana Santos, Anthony Hevia, Antigona Unldreaj, Arash Aghagol, Arezoo Abdollahi, Aycha Tammour, Azadeh HajiHosseini, Bahareh Behroozi, Benjamin Ajibade, Bharat Saxena, Carlos Muñoz Ferrandis, Daniel McDuff, Danish Contractor, David Lansky, Davis David, Douwe Kiela, Duong A. Nguyen, Edward Tan, Emi Baylor, Ezinwanne Ozoani, Fatima Mirza, Frankline Ononiwu, Habib Rezanejad, Hessie Jones, Indrani Bhattacharya, Irene Solaiman, Irina Sedenko, Isar Nejadgholi, Jesse Passmore, Josh Seltzer, Julio Bonis Sanz, Livia Dutra, Mairon Samagaio, Maraim Elbadri, Margot Mieskes, Marissa Gerchick, Martha Akinlolu, Michael McKenna, Mike Qiu, Muhammed Ghauri, Mykola Burynok, Nafis Abrar, Nazneen Rajani, Nour Elkott, Nour Fahmy, Olanrewaju Samuel, Ran An, Rasmus Kromann, Ryan Hao, Samira Alizadeh, Sarmad Shubber, Silas Wang, Sourav Roy, Sylvain Viguier, Thanh Le, Tobi Oyebade, Trieu Le, Yoyo Yang, Zach Nguyen, Abhinav Ramesh Kashyap, Alfredo Palasciano, Alison Callahan, Anima Shukla, Antonio Miranda-Escalada, Ayush Singh, Benjamin Beilharz, Bo Wang, Caio Brito, Chenxi Zhou, Chirag Jain, Chuxin Xu, Clémentine Fourrier, Daniel León Periñán, Daniel Molano, Dian Yu, Enrique Manjavacas, Fabio Barth, Florian Fuhrimann, Gabriel Altay, Giyaseddin Bayrak, Gully Burns, Helena U. Vrabec, Imane Bello, Ishani Dash, Jihyun Kang, John Giorgi, Jonas Golde, Jose David Posada, Karthik Rangasai Sivaraman, Lokesh Bulchandani, Lu Liu, Luisa Shinzato, Madeleine Hahn de Bykhovetz, Maiko Takeuchi, Marc Pàmies, Maria A Castillo, Marianna Nezhurina, Mario Sänger, Matthias Samwald, Michael Cullan, Michael Weinberg, Michiel De Wolf, Mina Mihaljcic, Minna Liu, Moritz Freidank, Myungsun Kang, Natasha Seelam, Nathan Dahlberg, Nicholas Michio Broad, Nikolaus Muellner, Pascale Fung, Patrick Haller, Ramya Chandrasekhar, Renata Eisenberg, Robert Martin, Rodrigo Canalli, Rosaline Su, Ruisi Su, Samuel Cahyawijaya, Samuele Garda, Shlok S Deshmukh, Shubhanshu Mishra, Sid Kiblawi, Simon Ott, Sinee Sang-aroonsiri, Srishti Kumar, Stefan Schweter, Sushil Bharati, Tanmay Laud, Théo Gigant, Tomoya Kainuma, Wojciech Kusa, Yanis Labrak, Yash Shailesh Bajaj, Yash Venkatraman, Yifan Xu, Yingxin Xu, Yu Xu, Zhe Tan, Zhongli Xie, Zifan Ye, Mathilde Bras, Younes Belkada, and Thomas Wolf. 2023. Bloom: A 176b-parameter open-access multilingual language model.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2022. ByT5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10:291–306.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

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*, pages 5397–5408, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Zhong Zhou, Lori S. Levin, David R. Mortensen, and Alexander H. Waibel. 2019. Using interlinear glosses as pivot in low-resource multilingual machine translation. *arXiv: Computation and Language*.

George Kingsley Zipf. 1945. The meaning-frequency relationship of words. *The Journal of General Psychology*, 33(2):251–256.

A Language Distribution

Figure 7 and Figure 8 display the number of examples per language and language family on a portion of our dataset.

B Grambank Typological Analysis

Along with number of languages, we would also like to measure whether the distribution of typological features in our dataset is reflective of the diversity of features in the world. We use the Grambank typological database (Skirgård et al., 2023) as a standard against which to judge the typological diversity of our dataset. Grambank covers over 2430 languages, with up to 195 features (e.g., "What is the order of numeral and noun in the NP?") per language. The values of the features comprise a vector for each language.

43% of our languages are found in Grambank, amounting to 72% coverage over all training instances. However, Grambank does not have complete feature vectors for all languages. Using the method described by Skirgård et al. (2023), we imputed missing feature values (9.7% of all features), resulting in complete feature vectors of size 161, as we only accept features that are defined for at least 64% of our dataset (as to balance dataset coverage as well as feature coverage).

We then create an average feature vector for our dataset by averaging the feature vectors of the the languages present in Grambank (weighting the average based on the number of training instances in each language) and compare this to the average of feature vectors for *all* languages in Grambank. We find a cosine similarity of 0.92 between the two vectors. In comparison, the language with the greatest similarity to the average Grambank vector in our data has a cosine similarity of 0.81. We report additional details of our methods and analysis below.

B.1 Imputation Details

We adapt the imputation procedure described in (Skirgård et al., 2023) and follow the steps below. Thresholds were chosen to maximize the language coverage while keeping the imputed values below 10%.

- Removed languages that had > 36% missing data out of the dataset
- Removed features that had >36% missing data among the remaining languages
- Binarized the multistate values
- Removed all but one dialect for each language according
- Imputed missing values with iterated random forest with MissForest[7]

B.2 Underrepresented Features

ID	Feature	Average Value
GB024b	Does the numeral always follow the noun in the NP?	0.468
GB193b	Are most adnominal property words placed after nouns?	0.536
GB025b	In the pragmatically unmarked order, does the adnominal demonstrative follow the noun?	0.425
GB118	Are there serial verb constructions?	0.406
GB130a	Is the order in intransitive clauses with a full nominal subject consistently SV?	0.382

Table 3: Most underrepresented Grambank features in training data with their average value (from imputed vectors).

In Table 3 we report the top five features for which the average representation in our training data is most distant from the expected value (as determined by averaging feature values across all languages in Grambank).

B.3 Averaged Feature Vector

For reference, we include the vector set of 161 average Grambank features over all training languages weighted by the number in Table 12.

C Training Hyperparameters

Training the GLOSSLM$_{\text{ALL}}$ and GLOSSLM$_{\text{UNSEG}}$ models used A6000 and A100 GPUs, and took around 5 days per run. We list the hyperparameters used in Table 4.

Parameter	Value
Optimizer	Adafactor
Initial LR	5e-5
Weight decay	0.01
Batch size	2
Gradient accumulation steps	64
Epochs	13

Table 4: Pretraining Hyperparameters

[7] https://rpubs.com/lmorgan95/MissForest

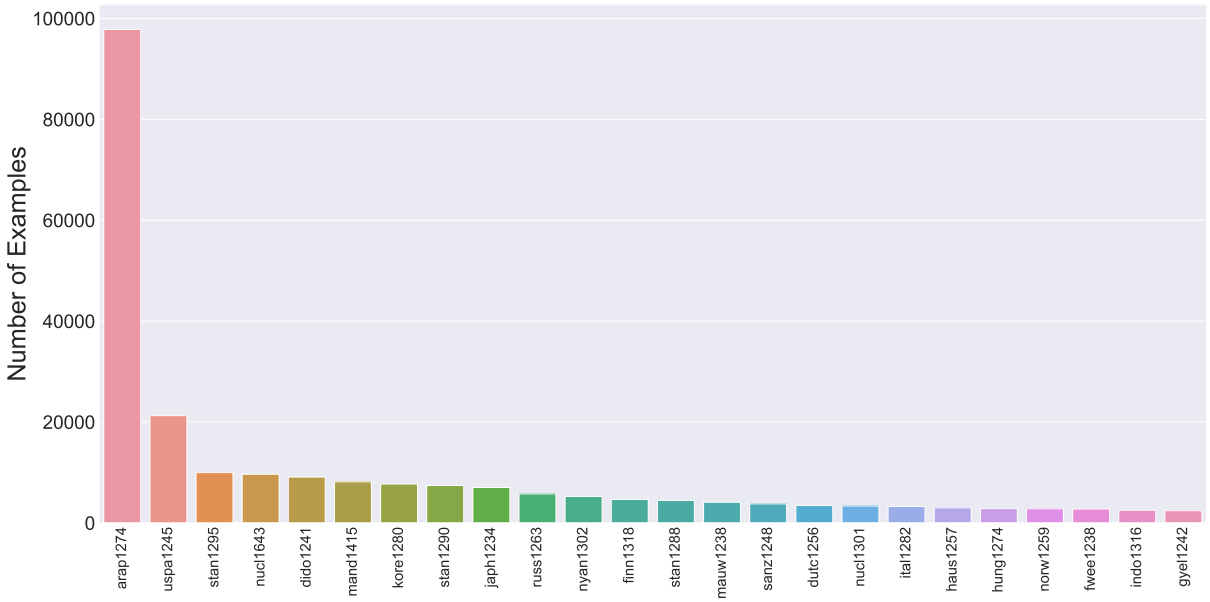

Figure 7: Counts per language. We only show languages with at least 2k samples present in the dataset. Arapaho (arap1274) is by far the most represented language in our data, followed by Uspanteko (uspa1245). Both languages are part of the SIGMORPHON Shared Task dataset.

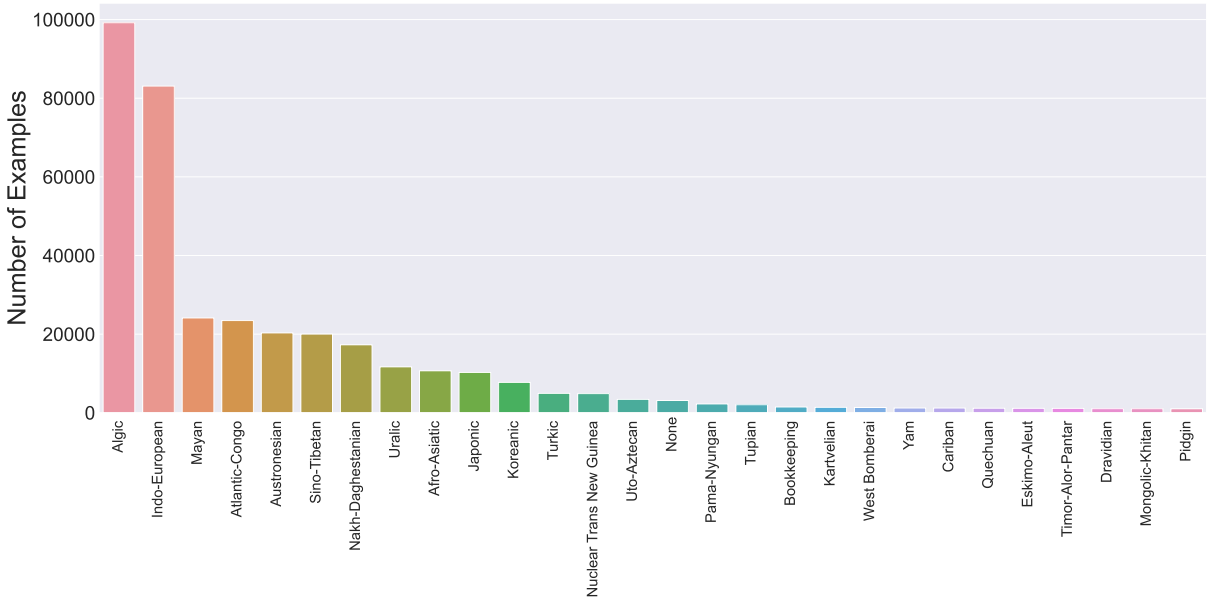

Figure 8: Counts per language family in our dataset. We only show language families with at least 1k samples present in the dataset. Language isolates and languages without recorded families in Glottolog (Hammarström et al., 2023) are categorized under "None".

Model	Morpheme accuracy / Word accuracy (Segmented)								
	In-domain				*Out-of-domain*				
	arp	**ddo**	**usp**	*Avg*	**git**	**lez**	**ntu**	**nyb**	*Avg*
TOP-CHOICE	83.2 / 74.0	78.5 / 64.4	79.7 / 72.9	80.5 / 70.4	51.1 / 29.7	62.2 / 54.4	78.4 / 68.1	72.5 / 63.8	66.1 / 54.0
TOKEN-CLASS	90.4 / 84.2	86.4 / 76.5	82.5 / 76.5	86.4 / 79.1	25.3 / 16.4	50.2 / 38.8	62.0 / 54.3	88.8 / 84.4	56.6 / 48.5
TÜ-CL	**90.7** / 84.6	91.2 / 85.1	84.5 / 78.5	88.8 / 82.7	50.2 / 26.6	**84.9** / **77.6**	**91.7** / 86.0	**91.4** / **87.9**	**79.6** / **69.5**
CRF	90.4 / 84.2	91.9 / 85.6	84.4 / 78.9	88.9 / 82.9	**52.4** / 33.6	84.7 / 77.5	91.1 / **86.6**	88.8 / 84.4	79.3 / 70.5
SMF	80.1 / 79.4	78.2 / 82.8	73.2 / 75.7	77.2 / 79.3	12.7 / 20.6	47.8 / 56.4	64.0 / 75.7	85.4 / 82.7	52.5 / 58.9
BYT5$_{\text{ALL}}$	88.7 / 83.2	**93.5** / **89.9**	86.3 / 82.7	89.5 / 85.3	2.2 / 3.6	72.5 / 69.7	83.4 / 82.2	90.7 / 89.2	62.2 / 61.2
GLOSSLM$_{\text{ALL, PRE}}$	89.3 / 84.2	91.7 / 88.3	84.1 / 81.0	88.4 / 84.5	3.6 / 9.1	3.6 / 1.8	4.9 / 9.8	2.9 / 3.0	3.8 / 5.9
GLOSSLM$_{\text{ALL, FT}}$	90.1 / **85.0**	92.8 / 89.3	**86.4** / **84.5**	**89.8** / **86.3**	28.9 / **34.9**	74.7 / 71.3	86.0 / 81.5	90.7 / 87.7	70.1 / 68.9

Model	Morpheme accuracy / Word accuracy (Unsegmented)								
	In-domain				*Out-of-domain*				
	arp	**ddo**	**usp**	*Avg*	**git**	**lez**	**ntu**	**nyb**	*Avg*
TOP-CHOICE	27.9 / 56.9	15.2 / 64.1	43.6 / 60.4	28.9 / 60.5	3.6 / 16.9	20.1 / 58.2	12.7 / 55.1	72.3 / 76.7	27.2 / 51.7
TOKEN-CLASS	43.6 / 69.9	51.2 / 74.3	57.2 / 72.1	50.7 / 72.1	8.54 / 16.9	40.7 / 45.5	19.4 / 48.2	14.2 / 5.96	20.7 / 29.1
TÜ-CL	77.8 / 77.5	74.1 / 80.4	70.0 / 73.4	74.0 / 77.1	**11.7** / 21.1	59.9 / **71.8**	56.2 / 78.0	85.2 / 85.0	53.3 / 64.0
BYT5$_{\text{UNSEG}}$	80.8 / 79.7	**84.2** / 87.4	**78.9** / **82.5**	81.3 / 83.2	0.1 / 0.3	42.2 / 53.4	53.7 / 71.0	**90.4** / **88.4**	46.6 / 53.3
GLOSSLM$_{\text{UNSEG, PRE}}$	79.8 / 79.2	77.5 / 82.8	76.8 / 80.8	78.0 / 80.9	2.3 / 3.5	1.5 / 1.3	4.1 / 9.6	1.6 / 2.9	2.4 / 4.3
GLOSSLM$_{\text{UNSEG, FT}}$	**82.1** / **81.5**	83.6 / 87.3	78.6 / 81.0	**81.4** / **83.3**	10.1 / **28.4**	57.3 / 64.9	62.8 / **78.9**	87.4 / 86.2	54.4 / **64.6**
GLOSSLM-NORM$_{\text{UNSEG, PRE}}$	79.6 / 80.0	79.6 / 83.2	74.8 / 76.6	78.0 / 79.9	2.2 / 7.8	2.6 / 1.8	2.9 / 9.7	1.0 / 2.5	2.2 / 5.45
GLOSSLM-NORM$_{\text{UNSEG, FT}}$	82.0 / **81.5**	**84.2** / **87.8**	76.4 / 79.2	80.8 / 82.8	9.3 / 16.4	**60.3** / 67.8	**63.4** / 76.6	90.0 / 87.6	**55.8** / 62.1

Table 5: Morpheme- and word-level accuracy of various systems on segmented (top) and unsegmented (bottom) text. Best performance per language in each setting the table is **bolded**. GLOSSLM$_{\text{ALL, PRE}}$ refers to performance using the pretrained GLOSSLM directly, while GLOSSLM$_{\text{ALL, FT}}$ refers to performance after fine-tuning the pretrained model on the specific language.

Model	chrF++ Score (Segmented)								
	In-domain				*Out-of-domain*				
	arp	**ddo**	**usp**	*Avg*	**git**	**lez**	**ntu**	**nyb**	*Avg*
TOP-CHOICE	75.0	71.9	71.4	72.8	33.7	69.7	74.5	62.2	60.0
TOKEN-CLASS	84.2	81.8	75.3	80.4	25.6	52.1	65.4	84.3	56.9
TÜ-CL	85.2	88.4	77.7	83.8	34.8	78.9	87.9	87.7	72.3
CRF	84.2	88.2	79.4	83.9	40.9	**79.2**	**88.4**	84.3	**73.2**
SMF	80.7	86.6	76.3	81.2	28.0	62.0	78.8	82.2	62.6
BYT5$_{\text{ALL}}$	84.2	**91.8**	84.6	86.9	9.07	74.9	85.1	**88.8**	64.5
GLOSSLM$_{\text{ALL, PRE}}$	85.2	90.8	83.1	86.4	21.5	14.1	18.2	8.8	15.7
GLOSSLM$_{\text{ALL, FT}}$	**86.3**	91.5	**85.9**	**87.9**	**43.1**	75.0	84.1	87.4	72.4

Model	chrF++ Score (Unsegmented)								
	In-domain				*Out-of-domain*				
	arp	**ddo**	**usp**	*Avg*	**git**	**lez**	**ntu**	**nyb**	*Avg*
TOP-CHOICE	44.0	63.5	55.1	54.2	8.4	51.6	40.5	74.0	43.6
TOKEN-CLASS	56.2	72.9	65.3	64.8	18.8	56.4	45.1	18.8	34.8
TÜ-CL	77.6	84.6	72.5	78.2	23.0	**71.5**	78.6	84.1	64.3
BYT5$_{\text{UNSEG}}$	80.7	**90.0**	**83.0**	84.6	7.6	59.6	77.0	**88.4**	58.2
GLOSSLM$_{\text{UNSEG, PRE}}$	80.5	86.8	81.0	82.8	19.4	13.3	16.3	8.1	14.3
GLOSSLM$_{\text{UNSEG, FT}}$	**82.9**	89.8	81.7	**84.8**	**34.9**	68.8	**80.7**	85.5	**67.5**
GLOSSLM-NORM$_{\text{UNSEG, PRE}}$	80.3	86.6	78.7	81.9	19.5	14.3	16.6	7.4	14.5
GLOSSLM-NORM$_{\text{UNSEG, FT}}$	82.6	89.9	81.1	84.5	29.1	70.9	79.0	87.5	66.6

Table 6: CHRF++ scores of various systems on segmented (top) and unsegmented (bottom) data. Best performance per language in each setting the table is **bolded**. GLOSSLM$_{\text{ALL, PRE}}$ refers to performance using the pretrained GLOSSLM directly, while GLOSSLM$_{\text{ALL, FT}}$ refers to performance after finetuning on the specific language.

D Full Results

We provide full results for accuracy and chrF++ scores in Table 5 and Table 6 . For our normalization experiments, we only trained and tested on unsegmented data for the target languages.

E In- vs out-of-vocabulary errors

	In-domain			*Out-of-domain*			
	arp	ddo	usp	git	lez	ntu	nyb
% OOV	30.0	15.6	20.0	78.1	27.3	27.6	8.42
IV	96.2	92.3	91.4	66.7	85.4	91.3	92.8
OOV	55.7	71.7	57.1	26.0	33.9	55.7	32.6
% OOV	30.0	18.7	21.4	80.5	25.5	28.9	9.27
IV	95.3	91.4	89.1	60.0	81.6	91.3	91.7
OOV	50.1	69.7	50.9	20.7	16.4	48.3	30.6

Table 7: Percent of words that are out-of-vocab in the test split for each language along with in- versus out-of-vocabulary accuracy at the word level. Top is the segmented setting (GLOSSLM$_{\text{ALL, FT}}$), bottom is unsegmented (GLOSSLM$_{\text{UNSEG, FT}}$).

Language	Morpheme %OOV
arp	3.6
ddo	41.2
usp	4.9
git	2.8
lez	1.1
ntu	0.5
nyb	5.3

Table 8: Percent of out-of-vocabulary morphemes in the test split.

Language	OOV Token Recall
arp	49.87
ddo	44.13
usp	44.89
git	58.21
lez	71.66
ntu	40.14

Table 9: Percent of lexical glosses present in the translation in the test split. Nyangbo examples do not include translations.

We report word-level accuracy for our finetuned GLOSSLM models, indicating whether the transcribed word is in- or out-of-vocabulary in Table 7, as well as the percent of OOV words in the test set. The OOV rate between segmented and unsegmented may vary slightly, as mappings between segmented and unsegmented forms are not necessarily one-to-one. We consider a word to be in-vocabulary if the form of the word in the transcription *and* its corresponding gold label in the gloss co-occur in the training data. We also include morpheme-OOV rates and statistics on lexical gloss overlap with translations in in Table 8 and Table 9, as reported in Ginn et al. (2023).

F Example Outputs

We include example outputs to show the errors discussed in §7.1 and §7.2.

Sample ID		
uspa1245_136	transcription	jaan
	translation	Esta bien.
	gold	bueno
	output	esta@bien
uspa1245_149	transcription	kond (ti') laj chaak
	translation	Cuando en su trabajo.
	gold	cuando ??? PREP trabajo
	output	cuándo??? PREP trabajo
lezg1247_1	transcription	вич- ни хъун- нва- й къвалах- ар я
	translation	and all this were real stories
	gold	reflxv- FOC - be- PERF- PST- word- PL was
	output	himself- FOC be- PERF- PTP real.story- PL was
lezg1247_22	transcription	ва гьада гьикъван гьа терези- ди- н стрелка пара хкаж хьа- нва- тIа гьам вилик кутуна- нва
	translation	she put first that person which is more valuable according to the position of scale arrow.
	gold	and then how that.the.same scale- DIR- GEN arrow very up happened- PERF- COND that before put- PERF
	output	that according.to that.the.same value- OBL- GEN position.in very hit be- PERF- COND that.the.same behind put- PERF
arap1274_991	transcription	neetotohoe
	translation	Take off your pants !
	gold	take.off.one's.pants
	output	take.off.pants
arap1274_1998	transcription	cee'iyo
	translation	payday .
	gold	payment
	output	pay.day
arap1274_1667	transcription	howouunonetiit hiniito'eibetiit biisiinowoot niihooku'oot beh- 'entou- '
	translation	pity , relationships , learning through observation , watching closely , it's all there .
	gold	pity.mercy relatedness learning.by.observing watching.along all- located.present 0S
	output	pity.mercy relationships learning.through.observation watching.along all- located.present- 0S

Table 10: Selected example outputs to illustrate errors by GLOSSLM$_{\text{ALL}}$ finetuned models.

Sample ID			MER
lezg1247_71	transcription	Акъадарда и пачагь ибуру ламрал , балкIандал акъадарда , ламрал акъадарда яда , цIайни ахъайда , им гьи хуьрей агъуз .	
	gold	mount.ENT this king these-ERG donkey.on horse-SPSS mount.ENT donkey.on mount.ENT or fire.and released.ENT he this village.out.of down	
	ByT5	mount.ENT-ENT this king these-ERG donkey-OBL-SPSS horse-OBL-SPSS mount.ENT-ENT donkey-OBL-SPSS mount.ENT-ENT was fire-FOC put.ENT-ENT he this village-INESS-SPSS down.ENT-ENT	0.585
	GlossLM	mount.ENT this king these-ERG donkey.on horse-SPSS mount.ENT donkey.on mount.ENT or fire.and mount.ENT he this village.out.of down	0.068
lezg1247_49	transcription	Хтана балкIан , « Гьаа », лагьана « гила чавай физ жеда » лагьана , « пачагьдин руш гьиз ».	
	gold	return-AOR horse Yes say-AOR now help-INELAT go-DAT be-ENT say-AOR king-DIR-GEN girl will.bring-DAT	
	ByT5	coming-AOC horse Yes say-AOC now king-DIR-GEN girl be-ENT say-AOC king-DIR-GEN girl be-ENT	0.398
	GlossLM	return-AOR horse Yes say-AOR now help-INELAT go-DAT be-ENT say-AOR king-DIR-GEN girl will.bring-DAT	0
lezg1247_27	transcription	АтIуз гана ибурун кьилерни .	
	gold	cutting.IMC-DAT give-AOR these-ERG-GEN chapter-PL-FOC	
	ByT5	then give-AOR these-ERG-GEN head-PL-FOC	0.380
	GlossLM	cutting.IMC-DAT give-AOR these-ERG-GEN chapter-PL-FOC	0
lezg1247_0	transcription	И гададини гьил вегьена са жуьт кьачуда .	
	gold	this boy-DIR-GEN-ERG hand threw-AOR one pair take-INESS	
	ByT5	this boy-DIR-GEN-ERG hand took-AOC one necklace take-ENT	0.327
	GlossLM	this boy-DIR-GEN-ERG hand threw-AOR one pair take-INESS	0
lezg1247_42	transcription	И рушни пачагь хьана гила башламишда вичин пачагьвализ .	
	gold	this girl-Q king happened-AOR now started.ENT-INESS himself-ERG-GEN reigned.ENT-ERG-DAT	
	ByT5	this girl-FOC king be-AOC now started.ENT-ENT himself-OBL-GEN reign.to-OBL-DAT	0.325
	GlossLM	this girl-Q king happened-AOR now started.ENT-INESS himself-ERG-GEN reigned.ENT-ERG-DAT	0

Table 11: Lezgi examples with the highest difference in MER between finetuned ByT5 and GlossLM outputs.

ID	Avg. Value	ID (cont.)	Avg. Value (cont.)	ID (cont.)	Avg. Value (cont.)
GB020	0.183	GB111	0.573	GB305	0.364
GB021	0.15	GB113	0.6	GB309	0.185
GB022	0.199	GB114	0.502	GB312	0.649
GB023	0.058	GB115	0.52	GB313	0.161
GB026	0.065	GB116	0.007	GB314	0.038
GB027	0.351	GB117	0.705	GB315	0.054
GB028	0.373	GB118	0.123	GB316	0.031
GB030	0.21	GB119	0.196	GB317	0.03
GB031	0.067	GB120	0.276	GB318	0.1
GB035	0.484	GB121	0.215	GB319	0
GB036	0.023	GB122	0.162	GB320	0.002
GB037	0.019	GB124	0.081	GB321	0.065
GB038	0.036	GB126	0.369	GB324	0.07
GB039	0.165	GB129	0.001	GB326	0.487
GB041	0.044	GB131	0.329	GB327	0.563
GB042	0.097	GB132	0.519	GB328	0.515
GB043	0.021	GB133	0.25	GB333	0.721
GB044	0.602	GB134	0.704	GB334	0.276
GB047	0.701	GB135	0.729	GB335	0.093
GB048	0.634	GB136	0.235	GB336	0.002
GB049	0.661	GB137	0.206	GB408	0.352
GB051	0.18	GB138	0.43	GB409	0.073
GB052	0.02	GB139	0.653	GB410	0.532
GB053	0.391	GB140	0.304	GB415	0.233
GB054	0.02	GB147	0.567	GB430	0.013
GB057	0.167	GB148	0.05	GB431	0.331
GB058	0.019	GB149	0.299	GB432	0.263
GB059	0.134	GB151	0.018	GB433	0.389
GB068	0.484	GB152	0.138	GB519	0.168
GB069	0.392	GB155	0.578	GB520	0.094
GB070	0.269	GB156	0.032	GB521	0.05
GB071	0.326	GB158	0.526	GB024a	0.664
GB072	0.538	GB159	0.176	GB024b	0.132
GB073	0.296	GB160	0.6	GB025a	0.678
GB074	0.543	GB165	0	GB025b	0.169
GB075	0.556	GB166	0.004	GB065a	0.644
GB079	0.51	GB167	0.036	GB065b	0.184
GB080	0.69	GB170	0.452	GB130a	0.469
GB081	0.032	GB171	0.179	GB130b	0.332
GB082	0.544	GB172	0.086	GB193a	0.626
GB083	0.647	GB177	0.272	GB193b	0.202
GB084	0.498	GB184	0.507		
GB086	0.653	GB185	0.292		
GB089	0.523	GB186	0.101		
GB090	0.146	GB192	0.098		
GB091	0.539	GB196	0.054		
GB092	0.106	GB197	0.041		
GB093	0.394	GB198	0.409		
GB094	0.143	GB257	0.519		
GB095	0.054	GB260	0.097		
GB096	0.018	GB262	0.095		
GB098	0.369	GB263	0.181		
GB099	0.113	GB264	0.062		
GB103	0.368	GB285	0.009		
GB104	0.286	GB286	0.154		
GB105	0.517	GB291	0.013		
GB107	0.521	GB297	0.059		
GB108	0.414	GB298	0.083		
GB109	0.107	GB299	0.314		
GB110	0.163	GB302	0.131		

Table 12: Grambank Feature Averages over Training Set