Robust Generalization Strategies for Morpheme Glossing in an Endangered Language Documentation Context

Michael Ginn and **Alexis Palmer**
University of Colorado
michael.ginn@colorado.edu and alexis.palmer@colorado.edu

Abstract

Generalization is of particular importance in resource-constrained settings, where the available training data may represent only a small fraction of the distribution of possible texts. We investigate the ability of morpheme labeling models to generalize by evaluating their performance on unseen genres of text, and we experiment with strategies for closing the gap between performance on in-distribution and out-of-distribution data. Specifically, we use weight decay optimization, output denoising, and iterative pseudo-labeling, and achieve a 2% improvement on a test set containing texts from unseen genres. All experiments are performed using texts written in the Mayan language Uspanteko.

1 Introduction

With over half of the world's languages endangered (Seifart et al., 2018), language documentation is one of several strategies for preservation. Traditionally, many documentation projects have aimed to create grammatical descriptions, dictionaries, and annotated text corpora, in the form of interlinear glossed text (IGT; see section 2.1). The annotated texts can be used in the creation of reference tools and pedagogical materials, as well as providing input data for downstream tasks such as machine translation (Zhou et al., 2019), morphological paradigm induction (Moeller et al., 2020), dependency parsing (Georgi et al., 2012), and other tasks (Georgi, 2016), making it particularly valuable for low-resource languages.

Annotation of large corpora can be time-consuming and monotonous, so there is a desire for systems to automatically produce IGT, annotating plain text with labels describing the part-of-speech, morphology, and syntax of each word in the corpus (Ginn et al., 2023). These systems can be used in conjunction with human annotators to create annotated corpora rapidly, ensuring consistency and reducing the amount of human effort required. Importantly, reducing annotation time also frees up language experts to work on other types of language preservation or revitalization activities.

However, generalization for automated annotation systems remains a critical problem. Pre-existing corpora of annotated text are often small, contain transcriptions of spoken language from a small number of distinct speakers, and focus on specific types of language such as story-telling and oration. Thus, systems trained on these corpora have difficulty generalizing to out-of-distribution (OOD) language, limiting their utility and robustness.

As acquiring additional annotated data is generally expensive and difficult, it is preferable to design models that generalize well to OOD data. In this work, we design models for one type of text annotation: labeling each morpheme in a text with its grammatical function. We envision these models being used alongside human annotators to provide suggestions and annotate text more quickly and consistently than by human labeling alone.

We examine three strategies to improve the robustness of these morpheme labeling models with limited data:

1. We optimize weight decay to improve generalization of large models.
2. We apply a separate denoiser model to improve performance on out-of-vocabulary inputs.
3. We apply self-supervised learning on unlabeled texts.

Our experiments evaluate model performance on texts of different genres than the texts in the training set, in order to investigate their ability to generalize to future, out-of-distribution texts. We find that these strategies achieve small performance improvements on in- and out-of-distribution texts,

Proceedings of the 1st GenBench Workshop on (Benchmarking) Generalisation in NLP, pages 89–98

and may be valuable for building more robust morpheme labeling models. Our code is available on GitHub.[1]

2 Background

2.1 Interlinear Glossed Text

In language documentation projects, annotated text typically uses a standardized format such as Interlinear Glossed Text (IGT) (Comrie et al., 2008), although the exact glossing conventions vary across projects. An example IGT sentence in Uspanteko is provided in 1.

(1) Ti- j- ya' -tq -a' juntiir
INC- E3S- *give* -PL -ENF *everything*
They give us everything
(Pixabaj et al., 2007)

The first line of the example is a **transcription** in the target language. Words may be transcribed as-is, or divided into morphemes (meaning-bearing units of language), as in the example.

The second line of the example gives a **gloss** for each morpheme. Glosses typically indicate either the translation of a morpheme or its grammatical function. For example, the *-tq-* morpheme is glossed as PL (plural). Stem morphemes, such as *ya'*, are glossed either with their translation (as here) or with a gloss indicating the stem type (such as VT for "transitive verb"). Our systems gloss stems using the latter approach.

The third line provides a translation of the sentence in a high-resource language, such as English.

Although there exist some large mixed-language corpora of IGT such as ODIN (Lewis and Xia, 2010) and IMTVault (Nordhoff and Krämer, 2022), the availability of IGT data is limited. For many languages, only small IGT corpora are available, and different corpora may (and do) use various annotation conventions. Depending on the wishes of the language community, such corpora may or may not be available for wider use or distribution.

2.2 Task

In this research, the task our systems address is to predict the gloss line of IGT given the transcription, segmented into morphemes. Each morpheme should be glossed with its grammatical function; to keep the output vocabulary small, we gloss stems with part-of-speech labels instead of translations. Using the example in item 1, the input to the system would be the sequence

"Ti j ya' tq a' [SEP] juntiir"

and the intended output would be

"INC E3S VT PL ENF [SEP] ADV"

where stems such as "ya'" and "juntiir" are glossed with the stem type, here VT for transitive verb and ADV for adverb.

2.3 Related Work

Existing scholarship has used a variety of approaches for automated gloss prediction, 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). Ginn and Palmer (2023) experiment with morphologically-inspired loss functions to improve low-resource glossing models. However, to our knowledge, there has been no evaluation or experimentation with generalization of these models.

One of the 2023 SIGMORPHON shared tasks involved creating models for automated gloss prediction (Ginn et al., 2023), with participant systems employing strategies such as leveraging the translation line for stem glossing (Okabe and Yvon, 2023), pretraining on large multilingual corpora (He et al., 2023), and straight-through gradient estimation (Girrbach, 2022).

Although the majority of machine learning research has traditionally evaluated models on in-distribution data, the ability to generalize to out-of-distribution data is desirable for natural language models (Linzen, 2020; Lake et al., 2017). This is particularly important for low-resource languages where collecting a wide distribution of data can be expensive or even infeasible.

3 Data & Methodology

3.1 Data

This work uses a corpus of IGT data for Uspanteko, a low-resource Mayan language, originally from the OKMA documentation project (Pixabaj et al., 2007) and adapted by Palmer et al. (2009). Morphemes are glossed with 68 different labels, plus a separator label. Each text was produced through recording speakers, transcribing text, and glossing

[1] https://github.com/michaelpginn/igt-glossing

with morpheme tags and translations. The corpus used includes 17 different speakers.

For this research, we experiment with generalization to unseen texts that represent different genres of text. This consideration is very practical for documentation projects, where the available training corpora are often the result of a single data collection project, and sometimes contain only one or two genres or registers of speech.

The Uspanteko corpus contains 27 texts in four different genres: stories, histories, personal anecdotes, and advice. We use the story and history texts as our in-distribution (ID) data, as we hypothesize that stories and histories have similar grammar and vocabulary. We use personal anecdotes and advice as our out-of-distribution (OOD) data. One intuitive difference between these sets is that stories and histories tend to talk about others, while an anecdote is about the speaker (and thus tends to use first-person voice) and advice is about the listener (second-person voice). There is only one instance where a document created by the same speaker appears in both the ID and OOD splits.

We randomly divide the ID data into training and evaluation sets and divide the OOD data into evaluation and final testing sets. The splits are listed in Table 1.

Set	Genre(s)	# Sentences
Training	Story, History	5049
Eval (ID)	Story, History	2128
Eval (OOD)	Personal, Advice	2128
Test (OOD)	Personal, Advice	2128

Table 1: Data splits, including in-distribution (ID) and out-of-distribution (OOD) data

To verify that these splits represent accurate distributions, we pretrained a masked language model on the training set (described in subsection 3.2) and calculated the perplexity for the ID and OOD eval sets.

Set	Perplexity
Eval (ID)	77.78
Eval (OOD)	94.03

Table 2: Perplexity of pretrained language model on data splits

Of course, genre and register only represent one form of out-of-distribution data. Data may also be out-of-distribution due to different speakers, dialects of a language, time period, and other factors.

All transcription data is segmented into morphemes. Thus, the task is to predict a gloss label for each morpheme in a sequence.

3.2 Pretraining

Existing pretrained models are rarely available for low-resource languages such as Uspanteko. Thus, we pretrain a new masked language model (MLM) on the training set before fine-tuning to the task at hand (on the same data set). We use a smaller variation of the RoBERTa architecture (Liu et al., 2019) to prevent over-fitting and reduce resources used. The model uses 3 hidden layers, hidden layers of size 100, and 5 attention heads, as in Gessler and Zeldes (2023), and we found in preliminary experiments that there is no significant difference in performance from a full-size RoBERTa model.

The model is pretrained using the parameters listed in Table 3. We employ a dynamic masking strategy (Liu et al., 2019) where 15% of tokens are masked, of which 80% use a *MASK* token, 10% use a random token, and 10% use the original token.

Parameter	Value
Optimizer	AdamW
β_1	0.9
β_2	0.999
ϵ	1E−8
Weight decay	0
Batch size	64
Gradient accumulation steps	3
Epochs	50
GPU	NVIDIA V100

Table 3: Training Hyperparameters
AdamW from Loshchilov and Hutter (2017b)

We refer to this pretrained model as USPMLM. For each experiment, USPMLM was fine-tuned on a token classification task. Because the words in the Uspanteko data are already segmented into morphemes, we are able to model this as a token classification task, predicting a gloss for each mor-

pheme. If segmentation were not available, we would have to model the problem with a sequence-to-sequence approach or use some strategy to predict morpheme segmentation. Still, in the token classification approach, the surrounding context for each morpheme is important to making high-quality predictions, and we cannot predict a gloss for each morpheme in a vacuum.

3.3 Evaluation

Models are evaluated on both the in-distribution and out-of-distribution evaluation sets. We follow the evaluation strategy used in the SIGMORPHON shared task, calculating the overall accuracy for every morpheme, ignoring word separators, and requiring glosses to be correctly aligned to morphemes.

4 Experiments

For a baseline model, we fine-tune USPMLM on the token classification task. Fine-tuning uses the same hyperparameters listed in Table 3. We also compare against a naïve strategy where we always select the most common gloss for a morpheme (based on the training data), as well as a strategy that selects a random gloss from the observed glosses for a morpheme in the training data.

We compare our baselines in Table 4.

Model	Acc. (Eval ID)	Acc. (Eval OOD)
Random	44.4	40.6
Most frequent	**85.0**	74.2
Neural	84.5	**74.6**

Table 4: Evaluation accuracy on in-distribution and out-of-distribution eval sets for baseline models

All strategies perform worse on the out-of-distribution data. The goal of the following experiments is to improve generalization of the model and thereby close the gap in performance for the ID and OOD evaluation sets. Though the neural model and the naïve model using the most frequent gloss perform similarly, we will conduct experiments with the neural model, which can be more readily manipulated to improve generalization.

4.1 Optimizing Weight Decay

Weight decay is important to avoiding overfitting and improving generalization (Loshchilov and Hutter, 2017a), helping reduce variance without sacrificing the representation power of larger models. We fine-tune USPMLM using six different values for weight decay; the results are listed in Table 5.

Weight Decay	Acc. (Eval ID)	Acc. (Eval OOD)
0 (Baseline)	84.5	74.6
0.01	84.2	73.7
0.1	84.3	74.0
0.5	**84.6**	74.8
0.75	**84.6**	**75.1**
1	84.5	74.4

Table 5: Evaluation accuracy for various weight decay values

We find that modifying the weight decay does not significantly affect the accuracy on ID data. However, for OOD data, the best-performing weight decay value of 0.75 achieves a 0.5 percentage point improvement over the baseline.

Generally, a weight decay of 0 or 0.01 is recommended, so it is interesting that a much larger value of 0.75 is successful in this case. These results could indicate that a more aggressive weight decay allows the model to better generalize to unseen documents, by reducing unnecessary weights and avoiding overfitting. However, the improvement is very small, and its possible that other techniques such as drop out are equally important for mitigating overfitting.

This result is likely heavily dependent on the model architecture, and the optimal weight decay value will vary from model to model. However, increasing weight decay beyond the typical recommendations seems to be an effective strategy.

4.2 Denoiser

4.2.1 Motivation

Generally, texts from out-of-distribution genres and registers will have more out-of-vocabulary (OOV) tokens in the input. This is the case in our data: the ID eval set has 4.3% unknown tokens and the OOD eval set has 9.6% unknown tokens.

Using the best-performing model from the previous section (weight decay of 0.75), we observe that a large portion of the error on the OOD eval set is a result of OOV morphemes in the input. The results of this analysis appear in Table 6.

	Eval (ID)	Eval (OOD)
# OOV Tokens	527	1322
# OOV Tokens Incor.	376	854
Total Incor.	1910	3447
Total Tokens	12388	13818
# OOV Incor. / Total Incor.	19.7%	24.8%
# OOV Incor. / Total Tokens	3.0%	6.2%

Table 6: Analysis of the error due to out-of-vocabulary (OOV) tokens in the in-distribution (ID) and out-of-distribution (OOD) eval sets

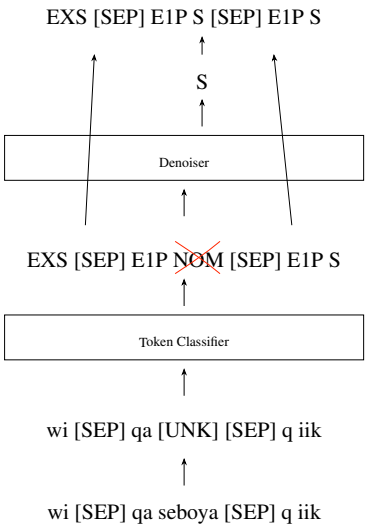

Figure 1: The denoising process. The morpheme "seboya" is OOV, and the token classifier makes an incorrect prediction. However, the denoiser uses observed label sequences to recover the correct gloss, which is substituted into the final prediction.

OOV tokens contributed 6.2 percentage points to the total error for the OOD data, and only 3.0 points for the ID data. Currently, the best model produces 15.4% error on the ID data and 24.9% error on the OOD, with a discrepancy of 9.5 percentage points. Thus, we observe that by reducing the error on OOV tokens, we can decrease a portion of this discrepancy.

4.2.2 Method

In many languages, morphological patterns are highly regular and structured, and some classes of morphemes (such as agreement morphology) may co-occur in fairly regular ways. We explore the potential of exploiting this fact to make better predictions on unknown morphemes using the other, known morphemes in the sentence. We train a **denoiser** language model on the gloss sequences in the training set. Then, we use this language model to predict gloss labels for OOV tokens, using the predicted glosses from the token classification model as the input to the denoiser (Figure 1).

The denoiser model, USPDENOISE, uses the same MLM architecture and training strategy as USPMLM. The model is trained with the hyperparameters in Table 3, except using a weight decay of 0.01 and 100 epochs.

For inference, we select the examples containing unknown morpheme tokens, and run USPDENOISE on the output of the fine-tuned token classification model. Then, we replace the prediction for each OOV morpheme with the prediction from the denoiser. We also experiment with masking the target tokens with the MASK token. We compare with the best-performing model from the previous section in Table 7.

Model	Acc. (Eval ID)	Acc. (Eval OOD)
No denoiser	84.6	75.1
Denoised (masked)	**84.7**	74.9
Denoised (no mask)	**84.7**	**75.3**

Table 7: Evaluation accuracy for denoiser strategies

The model using the denoiser without masking tokens shows the best performance, although the improvement is small. In this case, it evidently is difficult to recover the correct token from the surrounding contexts. However, this strategy could still be effective in cases where there are many OOV morphemes or the language is very regular.

4.3 Self-Supervision

4.3.1 Motivation

Perhaps the most effective way to improve performance on OOD data would simply be to train on OOD data, but in our example scenario this is not feasible. However, we can employ **iterative pseudo-labeling**, a form of self-supervised learning, to re-train the model using the labels predicted by a prior model (Chapelle et al., 2009). Iterative pseudo-labeling has been employed in low-resource speech recognition, where additional labeled data is similarly difficult to obtain (Kahn et al., 2020).

In the context of generalization, iterative pseudo-labeling can help adapt the model to the particular target distribution by re-training the model on predictions for the out-of-domain data. In this way, we can expose the model to the sort of contexts seen in the OOD data without needing additional labeling; retraining the model can also help when the target distribution uses different labeling conventions than the training set.

4.3.2 Method

Silovsky et al. (2023) uses iterative pseudo-labeling to improve performance for low-resource automated speech recognition (ASR) models. We follow their method, described here, hypothesizing that the improvements will be similar for glossing models.

First, we run predictions for our OOD eval set using the best-performing model from the previous section, with a weight decay of 0.75 and denoising. For each sentence, we compute a model confidence value by taking the softmax of the output logits to get the probability value for the most likely gloss at each position and then averaging these probabilities over all glosses in the sequence. We use these confidence values to rank the predictions for every sentence and select some fraction of the predictions with the highest confidence; we experimented with using the top half, third, and quarter of predictions.[2] We pseudo-label these examples with the predicted glosses.

Next, we re-train the trained model using the original training set combined with the selected pseudo-labeled examples. Iterative pseudo-labeling can be run for many iterations if the predictions continue to improve. The results after iterative pseudo-labeling for one iteration, using different fractions of the predictions, are shown in Table 8.

We find that the iterative pseudo-labeled models outperform the previous model, with the model using one-quarter of the pseudo-labeled data performing best on the OOD data (with a small tradeoff in ID performance). It seems that selecting a smaller amount of higher-confidence data is more effective than using additional lower-confidence predictions.

Next, we run iterative pseudo-labeling for additional iterations, using the model trained on the top quarter of predictions. In each iteration, we again select the top quarter of predictions, and fine-tune the model. The results after several iterations are given in Table 9.

[2]The effectiveness of this approach depends on how well-calibrated the model is.

Pseudo-labelled fraction	Acc. (Eval ID)	Acc. (Eval OOD)
0	84.7	75.3
1/4	85.8	**76.3**
1/3	**85.9**	76.2
1/2	85.5	75.8

Table 8: Evaluation accuracy for models using pseudo-labeling with different fractions of the eval set

Iteration	Acc. (Eval ID)	Acc. (Eval OOD)
0	84.7	75.3
1	85.8	76.3
2	**86.5**	**76.9**
3	86.3	76.8

Table 9: Evaluation accuracy after additional iterations of pseudo-labeling

The second iteration continues to provide performance benefits, but the third iteration shows a small decrease in performance, so we stop iterating and select the model after 2 iterations. While pseudo-labeling initially provides benefits by exposing the model to the target domain, after some iterations the additional noise introduced has detrimental effects. Overall, iterative pseudo-labeling improves the ID accuracy by 1.5 and the OOD accuracy by 1.6 percentage points.

5 Results

Table 10 provides the performance on the held-out, OOD test set using the best model from each step. Each model builds on the previous, so the final model uses all three strategies described in the paper.

In each step, we use the best trained model from the previous step. We do not iterate pseudo-labeling on the test set, since the test set should have the same distribution as the OOD eval set.

Through weight decay optimization, denoising, and iterative pseudo-labeling, we are able to accomplish an improvement of 2 percentage points in performance on OOD data, with an 8.2% reduction in overall error.

Model	Acc. (Test OOD)
Baseline	75.5
WD 0.75	76.0
Denoised	76.3
Pseudo-labeled	**77.5**

Table 10: Accuracy on held-out test set after applying each technique

These techniques also improve performance on the in-distribution eval set, although by a smaller margin than the out-of-distribution eval set. This is desirable, as it narrows the gap between performance on in- and out-of-distribution data, resulting in more predictable model performance.

6 Discussion

Although the techniques used in this work do yield performance improvements, generalization in language documentation remains a difficult task, largely due to hard-to-overcome challenges such as unseen morphemes, labels for morphemes that do not appear in the training set, and ambiguity in labeling.

Weight decay optimization, like all forms of hyperparameter tuning, is highly situation-dependent and requires good evaluation. Generally, avoiding overfitting and minimizing variance is critical to generalization in documentation, where the training sets may represent only a small fraction of the distribution of possible texts.

Denoising is a promising strategy for making high-quality predictions on completely unknown morphemes, using the surrounding context. This approach may be particularly useful in a human-in-the-loop situation, where the denoiser provides several top guesses for an unknown morpheme, and a human annotator can select between the options, allowing for easier annotation and possibly active learning (Palmer et al., 2009). Denoising will likely show more robust performance for languages with highly structured and productive morphological systems and relationships such as agreement and regular word order.

Some aspects of Uspanteko morphology are productive and structured. For example, verbs can take multiple affixes, both prefixes and suffixes, and these occur in a predictable order, according to a morphological pattern. At the same time, the language also has relatively flexible classes of morphemes, allowing non-verbal stems to act as predicates (Coon, 2016), taking some of the same morphology as seen on verb stems. This flexibility may have decreased the utility of the denoising approach, as unseen stems appearing in verbal positions could be verbal or non-verbal morphemes, with no clear distinction.

Iterative pseudo-labeling similarly shows only a small improvement. In these experiments, the OOD texts still share fairly similar contexts and labeling strategies with the training set, as evidenced by the perplexity values. However, in a case where the unseen texts are more dissimilar to the training set, this strategy could be more effective at tuning the model to the particular target distribution.

7 Future Research

This work presents a preliminary exploration into generalization for documentation models, and much work remains to be done. Documentation data for even the most widely-spoken languages is limited, yet robust generalization from the training set is crucial for improving usability.

One promising approach for creating more robust documentation models is through cross-lingual transfer that utilizes the morphological similarities between languages. He et al. (2023) demonstrates that this approach can effect performance improvements on in-distribution data, and it would likely benefit out-of-distribution data as well.

Another technique for avoiding overfitting and improving generalization is ensuring models focus on linguistic information, relying less on semantic patterns that may lead to spurious generalizations. This could involve morphologically inspired loss functions, data augmentation using rule-based systems, or pretraining on other linguistic tasks.

8 Conclusion

In this work, we presented three strategies for improving generalization of interlinear glossed text generation models, which to our knowledge are novel approaches to the problem. We use weight decay optimization, denoising, and iterative pseudo-labeling, finding that iterative pseudo-labeling provides the greatest improvement in performance. Overall, our best model achieves a 2% improvement from the baseline on a test set representing texts of unseen genres. We also investigate the discrepancy in performance between in- and out-

of-distribution data, finding that out-of-vocabulary morphemes and differences in context are key sources of error. We hope these approaches can inspire future work in improving generalization for documentation models, which is difficult but critical to the usability of automated documentation systems in real-world projects.

9 Limitations

This research was conducted testing on a single language and corpus, and the effectiveness of each approach may vary with the language used. Additionally, this work focused on glossing morphemes, provided words have already been segmented into morphemes. This is often not the case for IGT data, and segmentation remains a difficult problem.

The experiments utilized a single model architecture for consistency, but other architectures might show different performance. We used a small transformer architecture due to the size of the training dataset; a deeper network might show different results.

We focused on experimenting with texts of unseen genres as our out-of-distribution data, but this is only one form of generalization. Other types of OOD data include data from other speakers or communities, dialects of a language, and data from different documentation projects.

10 Ethical Considerations

When working with projects that affect language communities, we should always strive to avoid a colonialist approach, and we should bear in mind that language data does not exist in a vacuum, but is the product of human experience (Bird, 2020). Documentation projects should never be undertaken without the consent and cooperation of the relevant language community.

Generalization is desirable in order to produce more valuable documentation systems, but it can also cause the homogenization of language, which can particularly affect speakers of less widely spoken dialects.

Training large transformer models requires a large amount of computation and thus incurs an unavoidable carbon cost (Bender et al., 2021), and thus we aimed to keep the architectures as small as possible. Minimizing the environmental impact of machine learning is a critical ongoing area of research.

11 Acknowledgements

We thank the anonymous reviewers for their useful suggestions and feedback, as well as the LECS Lab at the University of Colorado. This material is based upon work supported by the National Science Foundation under Grant No. 2149404, "CAREER: From One Language to Another". 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

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.

Steven Bird. 2020. Decolonising Speech and Language Technology. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3504–3519, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Olivier Chapelle, Bernhard Scholkopf, and Alexander Zien. 2009. Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]. *IEEE Transactions on Neural Networks*, 20(3):542–542.

Bernard Comrie, Martin Haspelmath, and Balthasar Bickel. 2008. The Leipzig Glossing Rules: Conventions for interlinear morpheme-by-morpheme glosses. *Department of Linguistics of the Max Planck Institute for Evolutionary Anthropology & the Department of Linguistics of the University of Leipzig. Retrieved January*, 28:2010.

Jessica Coon. 2016. Mayan Morphosyntax: Mayan Morphosyntax. *Language and Linguistics Compass*, 10(10):515–550.

Ryan Georgi, Fei Xia, and William D. Lewis. 2012. Improving Dependency Parsing with Interlinear Glossed Text and Syntactic Projection. In *International Conference on Computational Linguistics*.

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

Luke Gessler and Amir Zeldes. 2023. MicroBERT: Effective Training of Low-resource Monolingual BERTs through Parameter Reduction and Multitask Learning. ArXiv:2212.12510 [cs].

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. Taxonomic loss for morphological glossing of low-resource languages.

Leander Girrbach. 2022. SIGMORPHON 2022 shared task on morpheme segmentation submission description: Sequence labelling for word-level morpheme segmentation. In *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 124–130, Seattle, Washington. Association for Computational Linguistics.

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.

Dieuwke Hupkes, Mario Giulianelli, Verna Dankers, Mikel Artetxe, Yanai Elazar, Tiago Pimentel, Christos Christodoulopoulos, Karim Lasri, Naomi Saphra, Arabella Sinclair, et al. 2023. A taxonomy and review of generalization research in NLP. *Nature Machine Intelligence*, 5(10):1161–1174.

Jacob Kahn, Ann Lee, and Awni Hannun. 2020. Self-training for end-to-end speech recognition. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7084–7088. IEEE.

Brenden M. Lake, Tomer D. Ullman, Joshua B. Tenenbaum, and Samuel J. Gershman. 2017. Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40:e253.

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.

Tal Linzen. 2020. How can we accelerate progress towards human-like linguistic generalization? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5210–5217, Online. Association for Computational Linguistics.

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].

Ilya Loshchilov and Frank Hutter. 2017a. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Ilya Loshchilov and Frank Hutter. 2017b. Fixing weight decay regularization in adam. *CoRR*, abs/1711.05101.

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.

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.

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.

Shu Okabe and François Yvon. 2023. 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.

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.

Telma Can Pixabaj, Miguel Angel Vicente Méndez, Maria Vicente Méndez, and Oswaldo Ajcot Damián. 2007. Text collections in four mayan languages.

Archived in The Archive of the Indigenous Languages of Latin America.

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

Jan Silovsky, Liuhui Deng, Arturo Argueta, Tresi Arvizo, Roger Hsiao, Sasha Kuznietsov, Yiu-Chang Lin, Xiaoqiang Xiao, and Yuanyuan Zhang. 2023. Cross-lingual knowledge transfer and iterative pseudo-labeling for low-resource speech recognition with transducers.

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*.

A GenBench Evaluation Card

Motivation			
Practical ☐	*Cognitive*	*Intrinsic*	*Fairness*
Generalisation type			
Compositional	*Structural*	*Cross Task*	*Cross Language*
Cross Domain ☐	*Robustness* ☐		
Shift type			
Covariate ☐	*Label*	*Full*	*Assumed*
Shift source			
Naturally occuring ☐	*Partitioned natural*	*Generated shift*	*Fully generated*
Shift locus			
Train–test ☐	*Finetune train–test*	*Pretrain–train*	*Pretrain–test*

Figure 2: GenBench evaluation card as described in Hupkes et al. (2023)