Lightweight morpheme labeling in context: Using structured linguistic representations to support linguistic analysis for the language documentation context

Bhargav Shandilya and Alexis Palmer
University of Colorado Boulder
{bhargav.shandilya,alexis.palmer}@colorado.edu

Abstract

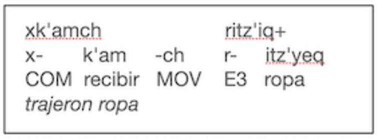
Linguistic analysis is a core task in the process of documenting, analyzing, and describing endangered and less-studied languages. In addition to providing insight into the properties of the language being studied, having tools to automatically label words in a language for grammatical category and morphological features can support a range of applications useful for language pedagogy and revitalization. At the same time, most modern NLP methods for these tasks require both large amounts of data in the language and compute costs well beyond the capacity of most research groups and language communities.

In this paper, we present a **gloss-to-gloss (g2g)** model for linguistic analysis (specifically, morphological analysis and part-of-speech tagging) that is lightweight in terms of both data requirements and computational expense.

The model is designed for the interlinear glossed text (IGT) format, in which we expect the source text of a sentence in a low-resource language, a translation of that sentence into a language of wider communication, and a detailed glossing of the morphological properties of each word in the sentence. We first produce silver standard parallel glossed data by automatically labeling the high-resource translation. The model then learns to transform source language morphological labels into output labels for the target language, mediated by a structured linguistic representation layer. We test the model on both low-resource and high-resource languages, and find that our simple CNN-based model achieves comparable performance to a state-of-the-art transformer-based model, at a fraction of the computational cost.

1 Introduction

Linguistic analysis is a core task in the documentation, analysis, and description of endangered and less-studied languages. One frequent goal of language documentation projects is to produce a corpus of **interlinear glossed texts**, or IGT (Figure 1 shows an example from the Mayan language Uspanteko). IGT can take many different forms, but canonically consists of the target language sentence, morphological segmentation of each word, glossing of each word with its stem translation and any relevant morphosyntactic features, and a translation into a language of wider communication.



xk'amch			ritz'iq+	
x-	k'am	-ch	r-	itz'yeq
COM	recibir	MOV	E3	ropa
trajeron ropa				

Figure 1: Example of IGT: Uspanteko (usp) sentence.

In addition to providing insight into the properties of the language being studied, the linguistic information in IGT can support a range of applications useful for language teaching and revitalization. Modern NLP methods typically require both large amounts of annotated data in the target language and compute resources beyond the capacity of most research groups and language communities. In this paper we address the task of **lightweight** morpheme labeling in context (McCarthy et al., 2019), developing a model which achieves reasonable accuracy with minimal requirements for *both* labeled data and computational expense.

Following previous work (Moeller and Hulden, 2021; Moeller et al., 2021; McMillan-Major, 2020; Zhao et al., 2020; Baldridge and Palmer, 2009, among others), **we aim to predict the parts of speech (POS) and morphosyntactic features for each word in the target sentence**, producing the third line of Figure 1, with stem translations replaced by POS labels. This model can produce a first-pass labeling for correction by human experts working on the documentation project, saving large amounts of time (as shown by Baldridge and Palmer, 2009) and freeing experts to work on more complex aspects of linguistic analysis.

To match the language documentation context,

Proceedings of the 20th SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology, pages 78–92
July 14, 2023 ©2023 Association for Computational Linguistics

where we often have a transcription and translation of the text before any other labeled data, we model morpheme labeling as a translation task. Specifically, the model should learn to transform labels for the high-resource translation into labels for the target language; hence the name **gloss-to-gloss (g2g)**.

For initial model development, we use data labeled in the UniMorph[1] format, so that we can test the model's performance on a range of languages. Next, we test the same model on Uspanteko data from a language documentation project (Pixabaj et al., 2007), which involves the steps of:

a) Converting the morpheme labels from the Uspanteko IGT into the UniMorph format, which includes mapping Uspanteko-specific labels into the UniMorph tag set;

b) Replacing stem translations (e.g. *ropa* (clothes)) with part-of-speech labels;

c) Translating the Spanish translations of the Uspanteko sentences into English, then automatically labeling the English text with UniMorph labels;

d) Using our g2g model to predict labels for the Uspanteko sentences.

For step (a), the expected UniMorph representation for the Uspanteko sentence above might be:

```
xk'amch      ritz'iq
V;PFV;ALL    N;ERG;3;PL
```

For example, the tag COM (completive aspect) from the Uspanteko IGT is mapped to the UniMorph label PFV (perfective aspect), and the tag E3 (ergative 3rd person plural) is converted to the UniMorph trio of ERG, 3, and PL.

Step (c) creates pseudo-parallel English data for the texts. For Figure 1, this step yields the following (noisy) morphological analysis:

```
[they]PRO;3;PL  [brought]V;PST
[clothes]N;PL
```

Even in this simple example, we see that the morphological information expressed in the two languages is similar but not identical, and the morphological features are distributed differently across the words. Our model additionally incorporates a layer that maps morpheme labels to their linguistic dimensions, following the dimensions defined by the UniMorph schema (Sylak-Glassman, 2016). Mapping to linguistic dimension is a first step toward incorporating linguistic knowledge for the task of morpheme glossing in context.

In step (d), we concatenate a vector of the English morpheme labels with static word embeddings for the English lexical items; this combined representation serves as input to the final classification layers, whose task is to produce the appropriate labels for the target language.

To keep computational demands low, we use a rather simple CNN-based architecture, and compare to a fine-tuned BERT (Devlin et al., 2019) model. On standard evaluations, the CNN model achieves performance comparable to the BERT model, at a fraction of the computational expense.

The contributions of this work are:

1. A lightweight (low computational expense, reasonable data requirements) model for morpheme labeling in context, with an architecture designed for a modified IGT (interlinear glossed text) format;

2. A simple structured linguistic representation in the form of linguistic dimensions, used to guide predictions;

3. Evaluation of the model on language documentation data (IGT) for the Mayan language Uspanteko, and additional evaluations on a range of high-resource languages.

We described related work in Section 2, our approach to data representation in Section 3, and the model architecture in Section 4. Section 5 describes results for the high-resource language development experiments, and Section 6 presents our core results on IGT for Uspanteko. We wrap up with discussion and conclusions.

2 Background and related work

One goal of this work is to develop time-saving tools for use in the language documentation context. Specifically, we aim to support the production of interlinear glossed text (IGT) with a lightweight model that can be run on a standard laptop, using whatever previously-produced IGT might be available for the target language.

2.1 Computational support for IGT

IGT is a standard format for representing rich linguistic information associated with text. It is a

[1] https://unimorph.github.io/

common representation in linguistic literature and a frequent product of language documentation and description projects.

At the same time, creating IGT is a time-consuming and expertise-demanding process, bringing together a collection of skilled tasks. Depending on the original data source, IGT production may require transcription and translation of recorded audio or video, as well as morphological segmentation and morphological analysis. An increasing amount of research effort has recently been devoted to finding low-resource solutions for each stage of the process, with work in transcription (for example, Adams et al., 2017; Wisniewski et al., 2020), translation (see Haddow et al. (2022) for a survey), and segmentation (Ruokolainen et al., 2013; Eskander et al., 2019; Mager et al., 2020, among others) tasks. Work on automatic morphological inflection for low-resource languages is also related, though it approaches the task from a different direction (Anastasopoulos and Neubig, 2019; Liu and Hulden, 2021; Muradoglu and Hulden, 2022; Wiemerslage et al., 2022, among others).

Representing IGT. Early computational efforts in this area focused on defining data formats for representing the complex relationships between the various tiers of IGT (Hughes et al., 2003, 2004; Palmer and Erk, 2007). The Xigt project (Goodman et al., 2015) improves upon and modernizes previous formats, offering an easily-serializable representation for IGT. In this study we take a different approach, extracting the morpheme labels from the IGT and clustering the labels for morphemes associated with a particular word into a UniMorph-style format (Batsuren et al., 2022). By using UniMorph, we depart from an important property of IGT: the direct and ordered association of labels with the morphemes they describe.

Morpheme glossing. The task of automatically producing IGT is the focus of a current (2023) shared task competition at SIGMORPHON.[2] Given a paired source text and translation, participants in the competition are asked to output, for each word, the appropriate stem translation and morpheme labels. Data are provided for seven different low-resource languages.

The earliest work on this task we are aware of (Baldridge and Palmer, 2009; Palmer et al., 2009, 2010) takes segmented data as input and outputs part-of-speech labels and morpheme labels, ignoring the stem translation part of the task. Samardžić et al. (2015) break the task down into two steps, starting with part-of-speech and morpheme labels and then filling in stem translations using dictionary resources, with predicted labels helping to disambiguate. Sequence labeling approaches, including Conditional Random Fields (CRFs), Hidden Markov Models, and Recurrent Neural Networks are explored by Barriga Martínez et al. (2021) for the Otomi language, and Moeller and Hulden (2018) consider both neural and non-neural sequence labeling approaches for several endangered languages. McMillan-Major (2020), who merge the outputs of two CRF models, one training on the source text, the other on the translation. Zhao et al. (2020) also leverage the translation signal for glossing.

In this work, we draw inspiration from earlier work in our focus on the morpheme labels (leaving aside the stem translation) and in our use of the translation to guide learning. We use a CNN to capture relationships between the source and target morpheme labels, combined with static word embeddings for the translated task to boost the semantic signal. The combination of these elements gives us a low-compute solution.

2.2 CNNs, and treating language as images

Convolutional Neural Networks (CNNs) have been used to some degree in NLP for static classification tasks and to capture latent structures in text. Before attention-based models became the standard approach to sequential prediction, CNNs were shown to achieve results that were comparable to other traditional language models such as RNNs and LSTMs. Pham et al. (2016) show that CNNs can be effective for dynamic sequence prediction tasks where both local and long-range dependency information needs to be captured. Their CNN model for statistical language modeling has a perplexity score comparable to popular RNN-based approaches.

A radically different approach to image-driven NLP is taken by Rust et al. (2023) to overcome vocabulary bottlenecks in languages. Their encoder approach (PIXEL) renders text as images and models orthographic similarity between languages. Although their approach does match BERT's performance on syntactic and semantic language tasks, PIXEL proves to be a more robust option for noisy

[2] https://github.com/sigmorphon/2023GlossingST

In 1923 she became a member of the Lägerdorf ADGB action committee. `[adp, num, 3;fem;nom;pro;sg, pst;ind;fin;v, det;indf, n;sg, adp, det;def, n;sg, sg;propn, n;sg, n;sg, _]`
1923 wurde sie Mitglied des Lägerdorfer ADGB - Aktionsausschusses. `[num, ind;3;v;sg;pst;fin;pass, 3;fem;sg;pro;nom, n;neut;sg;nom, gen;sg;def;det;masc, propn, sg;gen;masc;propn, _, n;sg;gen;masc, _]`

Table 1: Example of fully-prepared pseudo-parallel data. The source text is automatically-translated and glossed English; the target text is German.

text inputs. Kim et al. (2015) use a CNN coupled with an LSTM at the character level to perform language modeling. Although their model has 60% fewer parameters than popular LSTM architectures of the time, it outperforms word-level and morpheme-level LSTM baselines. Our work differs from this approach in that we encode both word order and morpheme-level information in two dimensions instead of using character-level representations.

3 Data and its representation

For model development, we start with data from the 2019 SIGMORPHON shared task on morphological analysis in context.[3] Once the model has been developed and tested, we apply it to a true low-resource language (Section 6.)

The shared task data is a collection of datasets of varying sizes, from 68 different languages and/or varieties, with sentence level morphological analysis in the UniMorph (Batsuren et al., 2022; McCarthy et al., 2020) style. Here we report results for 9 languages, selected for diversity of morphological systems. For each language, we select the first 10,000 sentences from the corpus and use a train/dev/test split of 60/20/20.

3.1 UniMorph data

The UniMorph (Universal Morphological Feature) schema is a set of morphological feature labels. This set of labels is intended to serve as an interlingua for annotation of (mostly) inflectional morphology, providing a universal schema into which any tag set can be mapped. The data consists of sentences, with lemmas and morphological labels assigned to each word within the sentence.

Pseudo-parallel data. Recall that our model treats morphological analysis as a translation task, "translating" the source-side labels into labels for the target-side sentence, assuming semantic equivalence. The UniMorph-labeled texts described above are not parallel, and we are not aware of any parallel texts with UniMorph-style labels. Therefore, we produce pseudo-parallel data by automatically translating each dataset into English and then labeling the English sentences with a morphological labeler trained on English UniMorph data. An instance of the fully prepared source data appears in Table 1.[4] We use the Google Translate API[5] to back-translate target text to English, our choice of high-resource anchor (source) language. We train a 64-unit BiLSTM model (Figure 8) with categorical cross-entropy loss to generate morphological labels for each word in the source language. We also trained a GRU model for the same purpose, but found that the BiLSTM is superior in terms of F1, as shown in Table 2.

3.2 Linguistic dimensions (LDs)

UniMorph's more than 200 individual labels are grouped into 23 linguistic dimensions, ranging from Aktionsart to voice, and including domains such as information structure and politeness (see Sylak-Glassman for details). For example, the marker PFV on the Uspanteko verb indicates completive aspect and can be mapped to the dimension of ASPECT. The marker PST on the Spanish verb indicates past tense, mapped to the linguistic dimension TENSE. We use the UniMorph linguistic dimensions in our model.

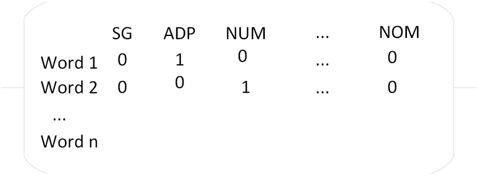
Model	Loss	Accuracy	F1
Bidirectional LSTM	0.111	0.969	0.922
GRU	0.126	0.963	0.876

Table 2: Performance of English glossing models.

[3] https://sigmorphon.github.io/sharedtasks/2019/task2/

[4] Further preprocessing involves removal of punctuation and conversion to all-lowercase letters.

[5] https://cloud.google.com/translate/docs



	SG	ADP	NUM	...	NOM
Word 1	0	1	0	...	0
Word 2	0	0	1	...	0
...					
Word n					

Figure 2: Multi-hot encoding for morpheme labels.

3.3 Structured representation for morphological features

Prior to encoding, the dataset consists of tokenized sentence and gloss pairs for the source and target languages. Each word in a sentence is naturally associated with one or more morphological features. This presents a multi-class encoding problem that is solved by using a categorical heat-map representation (Section 4), in which each column represents a single label (morphological feature or part-of-speech), and each row represents one word in the sentence. Considering the first three words in the sentence shown in Table 1, the encoding would be: *[In]->[adp], [1923]->[num], [she]->[3,fem,nom,pro,sg]*. The input to the model is a full 2-D binary representation where the column headers are the set of all possible individual morpheme labels and the rows consist of all words, with additional padding to standardize the input format. An example of the binary multi-hot encoding is shown in Figure 2. The gloss labels are then mapped to their linguistic dimensions (LDs).

4 Gloss-to-gloss (g2g) model

Figure 3 shows the architecture of the g2g system, and Figure 4 schematizes the model's workflow for one sample input sentence. Gloss labels for both source and target text are mapped to their linguistic dimensions (LDs) and encoded as heat maps, transforming the problem of glossing to an image-to-image prediction problem. The CNN generates heat maps with expectations over output gloss labels. These heat maps can be seen be seen as binary images, or alternately as sparse tensors. The heat maps, concatenated with pre-trained word2vec embeddings for the source language text, serve as inputs to shallow three-layer network for final labeling. The model's output prediction is a 2-D tensor of the same dimensions containing values that represent the probability of each morpheme label for each word in the sentence. We do not perform extensive parameter search, instead adopting standard settings (Appendix B).

4.1 Motivation

We assume parallel meaning between the source and target language texts. We also expect variability in how that meaning is expressed. **Translational divergence** can include challenges like differences in the structures employed by the two languages, differences in the morphological systems and their inventories, and variation with respect to what types of grammatical resources the languages use to convey the intended meaning. From a linguistic perspective, it is optimistic to expect a g2g approach to yield accurate target language glosses.

At the same time, we know that there are regularities to these divergences, and we expect our model to learn some of these mappings. To boost performance, we use word embeddings to capture meaning; these embeddings may also encode some information about morphology (Schwartz et al., 2022; Avraham and Goldberg, 2017; Soricut and Och, 2015). We use LDs to abstract away from particular labels into linguistic categories, and we use an extra probabilistic component (Section 4.3) to decide when to stop predicting labels.

4.2 Morphological representation and training

Before the morphological data is fed into the primary CNN model, we prime the representation with established classes/dimensions. For instance, we classify labels such as `ACC`, `NOM`, and `DAT` into the `CASE` category. This mapping reduces the count of label types by 60% on average for all source-target language pairs, consequently improving model accuracy and preventing mistakes that pertain to multiple category labels being predicted for the same word.

The heat map representation allows us to transform the sequence learning problem into a 2D image-based learning problem. The input is a binary image and the model output is a heat map that represents the probability values for the possible morpheme labels for each word in the sentence. Using this encoding format, we can create lightweight CNN models that can take inputs of any arbitrary padding (rows) and morphological feature (column) size. Due to the relatively small number of parameters, we can train a unique model for each language pair.

The input heat map images are fed to a standard convolutional neural network (CNN).[6] We obtain

[6] Further model details in Appendix B.

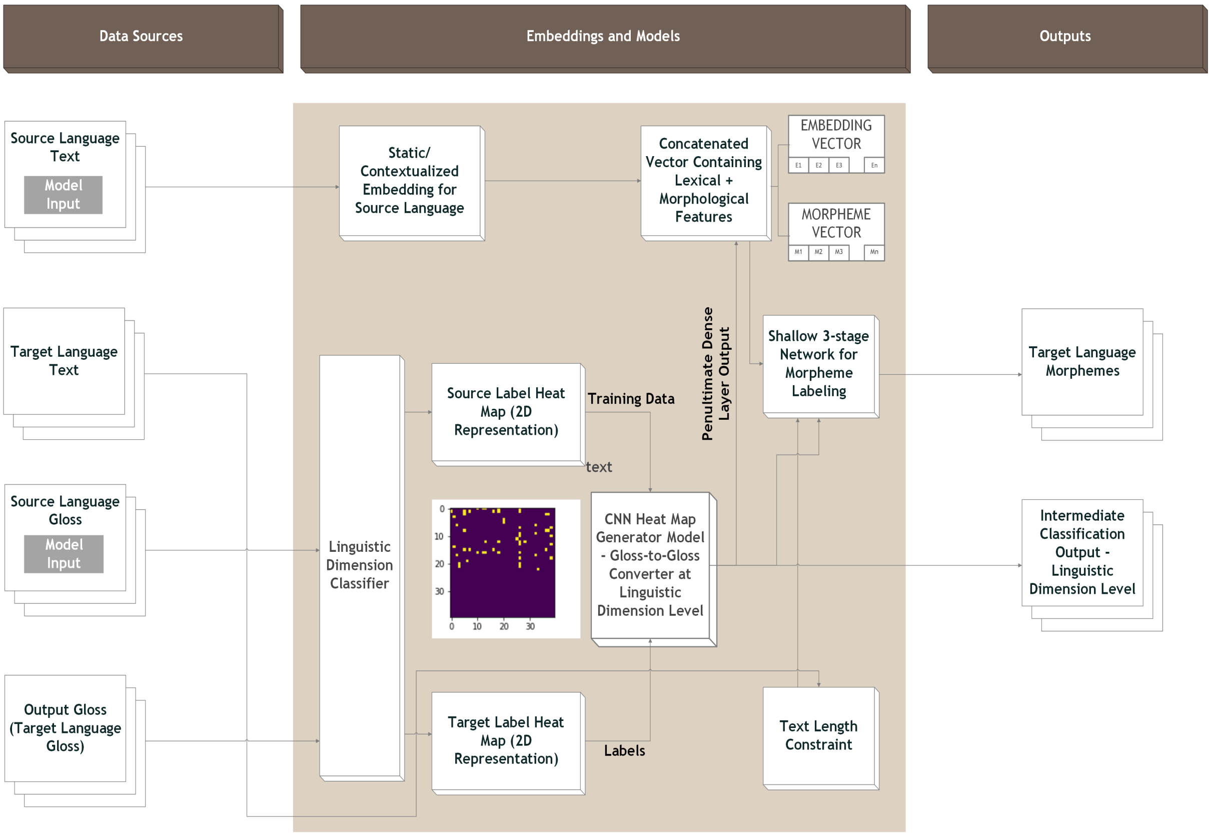

Data Sources
Embeddings and Models
Outputs
Source Language Text
Model Input
Static/ Contextualized Embedding for Source Language
Concatenated Vector Containing Lexical + Morphological Features
EMBEDDING VECTOR
E1 E2 E3 En
MORPHEME VECTOR
M1 M2 M3 Mn
Target Language Text
Penultimate Dense Layer Output
Shallow 3-stage Network for Morpheme Labeling
Target Language Morphemes
Source Label Heat Map (2D Representation)
Training Data
text
Source Language Gloss
Model Input
Linguistic Dimension Classifier
CNN Heat Map Generator Model - Gloss-to-Gloss Converter at Linguistic Dimension Level
Intermediate Classification Output - Linguistic Dimension Level
0
10
20
30
Output Gloss (Target Language Gloss)
Target Label Heat Map (2D Representation)
Labels
Text Length Constraint

Figure 3: Architecture of g2g system.

a discretized output of linguistic dimension predictions by setting a threshold and assigning hard categories to each cell (0 or 1). The threshold is considered a pipeline parameter for each language pair and is set by performing an optimal parameter search that maximizes the F1 score post-facto. The threshold for German, for instance, is set at 0.35.

4.3 Adding lexical information and probabilistic length modeling

To capture lexical semantics, we concatenate w2v embeddings (Mikolov et al., 2013) for the source side words with the penultimate dense layer of the CNN. We then train a shallow network with 3 dense layers on the concatenated vectors, outputting a flattened version of the heat map of target glosses. This one-dimensional representation is then transformed back into a 2D representation and decoded to obtain the target language gloss.

One challenge of the heat map approach is uncertainty about when to stop predicting labels. Sentence lengths vary, but the model always predicts a standard 40-word heat map. To address this issue, we use the sentence length of the target text (without lexical information; Zhao et al., 2020 use a similar approach) and a probabilistic model that determines the likelihood of a combination of morpheme labels occurring together and drops low-probability combinations (such as PRPN;PL (plural proper noun) for English) from the output heat map until the number of rows matches the number of words in the target language sentence. The selection is based on the joint probabilities of co-occurring morphological labels drawn from a likelihood lookup table constructed using the frequency of various possible morphological combinations in our training sets.

4.4 Training an LLM for morphological labeling - BERT

Since there is no easily available baseline for parallel text glossing, we train a BERT model to act as a comparable computationally-expensive baseline. Pre-trained cased weights are used since our common source language for all target languages is English. The possible gloss combinations in the target language form their own separate vocabulary and are together treated as a language of their own. The problem is reduced to a standard translation problem where English is the source and the gloss

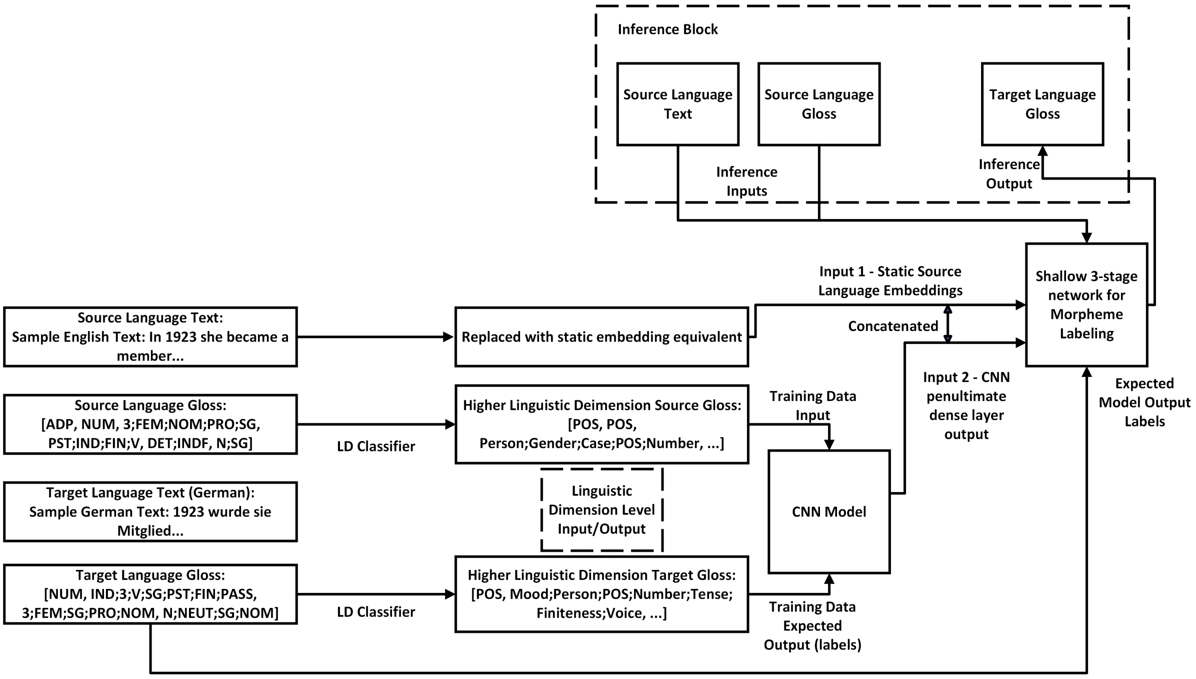

Inference Block
Source Language Text
Source Language Gloss
Target Language Gloss
Inference Inputs
Inference Output
Input 1 - Static Source Language Embeddings
Concatenated
Shallow 3-stage network for Morpheme Labeling
Source Language Text: Sample English Text: In 1923 she became a member...
Replaced with static embedding equivalent
Input 2 - CNN penultimate dense layer output
Expected Model Output Labels
Source Language Gloss: [ADP, NUM, 3;FEM;NOM;PRO;SG, PST;IND;FIN;V, DET;INDF, N;SG]
LD Classifier
Higher Linguistic Deimension Source Gloss: [POS, POS, Person;Gender;Case;POS;Number, ...]
Training Data Input
Target Language Text (German): Sample German Text: 1923 wurde sie Mitglied...
Linguistic Dimension Level Input/Output
CNN Model
Target Language Gloss: [NUM, IND;3;V;SG;PST;FIN;PASS, 3;FEM;SG;PRO;NOM, N;NEUT;SG;NOM]
LD Classifier
Higher Linguistic Dimension Target Gloss: [POS, Mood;Person;POS;Number;Tense; Finiteness;Voice, ...]
Training Data Expected Output (labels)

Figure 4: Workflow of g2g system for the sample input from Table 1.

	morpheme F1		LD F1		POS acc.	
	BERT	CNN	BERT	CNN	BERT	CNN
Basque	0.75	0.69	0.87	0.86	0.91	0.86
Finnish	0.81	0.77	0.86	0.84	0.92	0.84
French	0.81	0.83	0.85	0.89	0.95	0.91
German	0.78	0.75	0.91	0.79	0.88	0.83
Italian	0.79	0.75	0.84	0.82	0.94	0.91
Russian	0.82	0.73	0.89	0.78	0.88	0.84
Spanish	0.73	0.65	0.88	0.87	0.96	0.92
Turkish	0.78	0.66	0.79	0.78	0.87	0.80
English	0.84	0.82	0.95	0.89	0.95	0.89

Table 3: Performance of CNN and BERT models across languages. Morpheme-level=F1 over all labels, LD-level=F1 over linguistic dimension categories, POS=accuracy. F1 is computed following SIGMORPHON 2019 shared task metric.

vocabulary is the target. Concatenating the source language morphology vector with the BERT embedding did not significantly affect the output, so we use only contextual and positional embeddings to fine-tune the model, with a separate fine-tuned model for each source-target language pair.

5 Experiments, results, and discussion

To evaluate performance of the model, we test it on nine different language pairs (see Table 3). For all non-English languages, we back-translate to English. For English, our source language is German.

As a baseline, we fine-tune a pre-trained BERT model for the morpheme labeling task, using the same data and splits, but in a standard supervised learning set-up (i.e. no parallel data and no LDs).

The CNN model experiments were run on a 2.6 GHz Intel(R) Core(TM) CPU, taking an average of **8.5 minutes** to train. The BERT baseline experiments were run on a multi-GPU cluster, taking an average of **3.5 hours** to train.

Evaluation. Table 3 shows results for both models across all languages, with accuracy for POS labels and, for morpheme labels and linguistic dimensions, F1 as defined for the SIGMORPHON 2019 shared task: true positives are the set intersection of the gold and predicted labels for a word, and false positives are labels in the predicted set but not the gold.[7] All measures are computed at the heatmap level for each row (sentence) and averaged over the full dataset.

Results. Some patterns hold across most language pairs. For the most part, the CNN does not quite match the performance of the transformer. Crucially, though, the CNN trains on a single laptop in under 10 minutes, where the transformer needs a compute cluster and multiple hours to train. The CNN performance is generally within 5 percentage points of the BERT model, and this may be

[7] NOTE: although we use the same data and evaluation as the shared task, our CNN results are not directly comparable, because, unlike almost all participating teams, we do not use target language lexemes or labels as training input.

an acceptable performance in most documentation contexts - an empirical question for future work.

The POS score represents the proportion of the part of speech predictions that were correct. Because there are latent associations between POS category and morpheme labels (for example, it would be highly unusual to see aspectual features marked on nouns), the POS score should be directly proportional to the final F1 scores that we obtain for each language. This is reflected across our results. While both models struggle with Turkish and Russian, the CNN also performs poorly on Basque.

At the LD level, the model's performance is somewhat similar across the three Romance languages we considered (Spanish, French and Italian). While the CNN fails to perform better than the transformer in most scenarios, it is interesting to note that the CNN performs marginally better than the transformer on French. However, both models show a significant performance dip when it comes to Spanish.

The CNN's F1 dips to 0.66 for Turkish and 0.73 for Russian. This may be due to their high morphological complexity.

6 Applying the model to language documentation data

Finally, we apply our model to data from the Mayan language Uspanteko (Pixabaj et al., 2007), using the train/dev/test splits defined for the SIGMORPHON 2023 shared task: training on 21 texts (9774 sentences), and using one text each for dev and test (around 200 sentences each). The model's performance on language documentation data parallels the results for high-resource languages.

Experimental setup and data preparation Translations of the Uspanteko sentences are available in Spanish. To remain consistent, we translate the Spanish sentences to English using the Google Translate API.

It is important to note that this translation step adds compounding errors to the model's final gloss output.

IGT to UniMorph Mapping The labels used in the Uspanteko IGT belong to the glossing conventions selected by the language documentation project. The label set is particular to the linguistic properties of the language, and as such they make some different distinctions than those encoded by UniMorph. Some of the mappings between UniMorph and IGT are shown in Table 9. The custom mapping table that we built to convert IGT to UniMorph are available in our repository.[8]

Model	**F1**
BERT Linguistic Dimension Level	0.80
CNN Linguistic Dimension Level	0.76
BERT Morpheme Level	0.71
CNN Morpheme Level	0.63

Table 4: Performance of models on Uspanteko data.

Results and Discussion As seen in Table 4, the model performance on Uspanteko is comparable to its performance on morphologically complex high-resource languages like Turkish. This leads us to believe that our computationally efficient approach can indeed be used in the low-resource language documentation context to produce a first pass labeling, thus reducing the time an expert needs to spend on labeling. To better understand the system's errors, we show the label distribution for false positives output by the CNN model (Figure 5) at the level of linguistic dimension. 33% of these are the `unk` (unknown) label, which occurs when the model fails to make a confident prediction on the linguistic dimension. These are precisely the cases where the human expert should intervene.

We note that the model is not over-predicting the LD of part-of-speech, despite the fact that 42% of gold labels in the test set are part-of-speech labels. Instead, the model makes more errors for the categories of case, person, and number. We expect that the prevalence of case errors comes from the fact that Uspanteko uses an ergative-absolutive case system, with patterning entirely different from the rather impoverished nominative-accusative case system of English. Uspanteko also uses a number of grammatical categories not present in English, such as directionals and relational nouns (Tyers and Henderson, 2021). Looking at particular parts of speech, the model does well on conjunctions (84% accuracy), and struggles with adverbs and adjectives. 24% of adverb predictions are confused with adjective tags and about 13% of adverbs and adjectives are labeled 'unknown'.

[8] https://github.com/bhargav-ns/G2G_Conversion

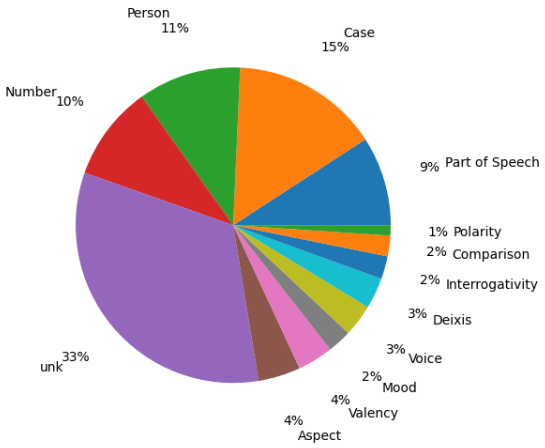


Figure 5: CNN: distribution of false positive predictions at the linguistic dimension level for Uspanteko.

7 Conclusion

We have presented the g2g model, a new architecture for morphological analysis that dramatically reduces compute time by modeling the task as, essentially, an image-to-image translation task. The model incorporates knowledge of linguistic categories by mapping labels to their linguistic dimensions, with the effect of narrowing the space of possible outputs. These strategies result in an enormous reduction of compute time and a system more suitable for use in low-resource scenarios than the large language models currently achieving top performance for this and similar tasks.

Model variants and future work. Working with language documentation data adds several layers of complexity. In this work, our model's output lacks the ordered association with individual morphemes typical of most IGT. We use an unordered set of labels to describe the morphological features of a word, as shown in Table 1.

In addition, there is wide variability in both the label sets and the glossing scheme across language documentation projects. One widely-used scheme is encoded in the Leipzig Glossing Rules (Bernard Comrie, 2008); Table 5 shows an example of a German phrase glossed according to Leipzig conventions. In future work we aim to produce outputs that mimic the glossing conventions used in the original data, including the order of the labels, the nature of the labels, and the glossing syntax.

The Sigmorphon 2023 shared task on interlinear glossing[9] hews closer to this goal. In this shared task, the source language text and target language text are used as inputs to obtain the target language glosses in Leipzig format. This is fundamentally different from our input format, as we attempt to obtain the target language glosses without the target language text. Instead, we use all the information available from the high-resource source language (text and glosses) as inputs to the model.

unser-	n	Väter-	n
our-	DAT.PL	father	DAT.PL
"To	our	fathers."	

Table 5: German phrase labeled using Leipzig Glossing Rules.

Our next step is to work directly with documentary linguists to evaluate whether and how such tools can be usefully deployed by field linguists and/or language community members. Another planned direction is to work on more sophisticated approaches to incorporating linguistic knowledge.

Limitations and ethical considerations. First, the system's performance is constrained by the use of automated systems to produce pseudo-parallel data. Errors in translation and morpheme labeling on the high-resource side propagate to the output and cause mistakes in target side labeling. We have not yet performed the extensive error analyses needed to understand how much error propagation might be affecting the system.

Second, we have not yet tested the system in an actual documentation project. When working on NLP with endangered and/or indigenous languages in mind, there is a clear risk of perpetuating existing oppression (Bird, 2020; Schwartz, 2022). We hope to avoid some of these harms by using data from a wide range of non-threatened languages first, waiting to involve language community members and documentary linguists until we have a system with good enough results that we expect it could actually be helpful in real world contexts. We have already developed collaborations with several speakers of endangered languages and linguists working on documentation projects, and we look forward to continuing this work with their guidance and involvement.

Acknowledgements

We thank the anonymous reviewers for very useful suggestions and feedback. Thanks also to the CLASIC cohort, FOLTA lab members, and CompSem

[9] https://github.com/sigmorphon/2023glossingST

lab members at CU Boulder for helpful discussions and feedback. 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

Oliver Adams, Trevor Cohn, Graham Neubig, and Alexis Michaud. 2017. Phonemic transcription of low-resource tonal languages. In *Proceedings of the Australasian Language Technology Association Workshop 2017*, pages 53–60, Brisbane, Australia.

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

Oded Avraham and Yoav Goldberg. 2017. The interplay of semantics and morphology in word embeddings. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 422–426, Valencia, Spain. Association for Computational Linguistics.

Jason Baldridge and Alexis Palmer. 2009. How well does active learning *actually* work? Time-based evaluation of cost-reduction strategies for language documentation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 296–305, Singapore. Association for Computational Linguistics.

Diego Barriga Martínez, Victor Mijangos, and Ximena Gutierrez-Vasques. 2021. Automatic interlinear glossing for Otomi language. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 34–43, Online. Association for Computational Linguistics.

Khuyagbaatar Batsuren, Omer Goldman, Salam Khalifa, Nizar Habash, Witold Kieraś, Gábor Bella, Brian Leonard, Garrett Nicolai, Kyle Gorman, Yustinus Ghanggo Ate, Maria Ryskina, Sabrina Mielke, Elena Budianskaya, Charbel El-Khaissi, Tiago Pimentel, Michael Gasser, William Abbott Lane, Mohit Raj, Matt Coler, Jaime Rafael Montoya Samame, Delio Siticonatzi Camaiteri, Esaú Zumaeta Rojas, Didier López Francis, Arturo Oncevay, Juan López Bautista, Gema Celeste Silva Villegas, Lucas Torroba Hennigen, Adam Ek, David Guriel, Peter Dirix, Jean-Philippe Bernardy, Andrey Scherbakov, Aziyana Bayyr-ool, Antonios Anastasopoulos, Roberto Zariquiey, Karina Sheifer, Sofya Ganieva, Hilaria Cruz, Ritván Karahóǧa, Stella Markantonatou, George Pavlidis, Matvey Plugaryov, Elena Klyachko, Ali Salehi, Candy Angulo, Jatayu Baxi, Andrew Krizhanovsky, Natalia Krizhanovskaya, Elizabeth Salesky, Clara Vania, Sardana Ivanova, Jennifer White, Rowan Hall Maudslay, Josef Valvoda, Ran Zmigrod, Paula Czarnowska, Irene Nikkarinen, Aelita Salchak, Brijesh Bhatt, Christopher Straughn, Zoey Liu, Jonathan North Washington, Yuval Pinter, Duygu Ataman, Marcin Wolinski, Totok Suhardijanto, Anna Yablonskaya, Niklas Stoehr, Hossep Dolatian, Zahroh Nuriah, Shyam Ratan, Francis M. Tyers, Edoardo M. Ponti, Grant Aiton, Aryaman Arora, Richard J. Hatcher, Ritesh Kumar, Jeremiah Young, Daria Rodionova, Anastasia Yemelina, Taras Andrushko, Igor Marchenko, Polina Mashkovtseva, Alexandra Serova, Emily Prud’hommeaux, Maria Nepomniashchaya, Fausto Giunchiglia, Eleanor Chodroff, Mans Hulden, Miikka Silfverberg, Arya D. McCarthy, David Yarowsky, Ryan Cotterell, Reut Tsarfaty, and Ekaterina Vylomova. 2022. UniMorph 4.0: Universal Morphology. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 840–855, Marseille, France. European Language Resources Association.

B. Bickel Max Planck Institute for Evolutional Anthropology Bernard Comrie, M. Haspelmath. 2008. *The Leipzig Glossing Rules: Conventions for Interlinear Morpheme-by-morphene Glosses*.

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Ramy Eskander, Judith Klavans, and Smaranda Muresan. 2019. Unsupervised morphological segmentation for low-resource polysynthetic languages. In *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 189–195, Florence, Italy. Association for Computational Linguistics.

Michael Wayne Goodman, Joshua Crowgey, Fei Xia, and Emily M. Bender. 2015. Xigt: extensible interlinear glossed text for natural language processing. *Language Resources and Evaluation*, 49(2):455–485.

Barry Haddow, Rachel Bawden, Antonio Valerio Miceli Barone, Jindřich Helcl, and Alexandra Birch. 2022. Survey of low-resource machine translation. *Computational Linguistics*, 48(3):673–732.

Baden Hughes, Steven Bird, and Catherine Bow. 2003. Encoding and presenting interlinear text using XML technologies. In *Proceedings of the Australasian Language Technology Workshop 2003*, pages 61–69, Melbourne, Australia.

Baden Hughes, Catherine Bow, and Steven Bird. 2004. Functional requirements for an interlinear text editor. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Yoon Kim, Yacine Jernite, David Sontag, and Alexander M. Rush. 2015. Character-aware neural language models.

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

Manuel Mager, Özlem Çetinoğlu, and Katharina Kann. 2020. Tackling the low-resource challenge for canonical segmentation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5237–5250, Online. Association for Computational Linguistics.

Arya D. McCarthy, Christo Kirov, Matteo Grella, Amrit Nidhi, Patrick Xia, Kyle Gorman, Ekaterina Vylomova, Sabrina J. Mielke, Garrett Nicolai, Miikka Silfverberg, Timofey Arkhangelskiy, Nataly Krizhanovsky, Andrew Krizhanovsky, Elena Klyachko, Alexey Sorokin, John Mansfield, Valts Ernštreits, Yuval Pinter, Cassandra L. Jacobs, Ryan Cotterell, Mans Hulden, and David Yarowsky. 2020. UniMorph 3.0: Universal Morphology. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3922–3931, Marseille, France. European Language Resources Association.

Arya D. McCarthy, Ekaterina Vylomova, Shijie Wu, Chaitanya Malaviya, Lawrence Wolf-Sonkin, Garrett Nicolai, Christo Kirov, Miikka Silfverberg, Sabrina J. Mielke, Jeffrey Heinz, Ryan Cotterell, and Mans Hulden. 2019. The SIGMORPHON 2019 shared task: Morphological analysis in context and cross-lingual transfer for inflection. In *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 229–244, Florence, Italy. Association for Computational Linguistics.

Angelina McMillan-Major. 2020. Automating gloss generation in interlinear glossed text. In *Proceedings of the Society for Computation in Linguistics 2020*, pages 355–366, New York, New York. Association for Computational Linguistics.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, Atlanta, Georgia. Association for Computational Linguistics.

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 and Mans Hulden. 2021. Integrating automated segmentation and glossing into documentary and descriptive linguistics. In *Proceedings of the 4th Workshop on the Use of Computational Methods in the Study of Endangered Languages Volume 1 (Papers)*, pages 86–95, Online. Association for Computational Linguistics.

Sarah Moeller, Ling Liu, and Mans Hulden. 2021. To POS tag or not to POS tag: The impact of POS tags on morphological learning in low-resource settings. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 966–978, Online. Association for Computational Linguistics.

Saliha Muradoglu and Mans Hulden. 2022. Eeny, meeny, miny, moe. how to choose data for morphological inflection. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7294–7303, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Alexis Palmer and Katrin Erk. 2007. IGT-XML: An XML format for interlinearized glossed text. In *Proceedings of the Linguistic Annotation Workshop*, pages 176–183, Prague, Czech Republic. 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, Boulder, Colorado. Association for Computational Linguistics.

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.

Ngoc-Quan Pham, German Kruszewski, and Gemma Boleda. 2016. Convolutional neural network language models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1153–1162, Austin, Texas. 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.

Teemu Ruokolainen, Oskar Kohonen, Sami Virpioja, and Mikko Kurimo. 2013. Supervised morphological segmentation in a low-resource learning setting using conditional random fields. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 29–37, Sofia, Bulgaria. Association for Computational Linguistics.

Phillip Rust, Jonas F. Lotz, Emanuele Bugliarello, Elizabeth Salesky, Miryam de Lhoneux, and Desmond Elliott. 2023. Language modelling with pixels.

Tanja Samardžić, Robert Schikowski, and Sabine Stoll. 2015. Automatic interlinear glossing as two-level sequence classification. In *Proceedings of the 9th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*, pages 68–72, Beijing, China. 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.

Lane Schwartz, Coleman Haley, and Francis Tyers. 2022. How to encode arbitrarily complex morphology in word embeddings, no corpus needed. In *Proceedings of the first workshop on NLP applications to field linguistics*, pages 64–76, Gyeongju, Republic of Korea. International Conference on Computational Linguistics.

Radu Soricut and Franz Och. 2015. Unsupervised morphology induction using word embeddings. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1627–1637, Denver, Colorado. Association for Computational Linguistics.

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

Francis Tyers and Robert Henderson. 2021. A corpus of k'iche' annotated for morphosyntactic structure. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 10–20, Online. Association for Computational Linguistics.

Adam Wiemerslage, Miikka Silfverberg, Changbing Yang, Arya McCarthy, Garrett Nicolai, Eliana Colunga, and Katharina Kann. 2022. Morphological processing of low-resource languages: Where we are and what's next. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 988–1007, Dublin, Ireland. Association for Computational Linguistics.

Guillaume Wisniewski, Séverine Guillaume, and Alexis Michaud. 2020. Phonemic transcription of low-resource languages: To what extent can preprocessing be automated? In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 306–315, Marseille, France. European Language Resources association.

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.

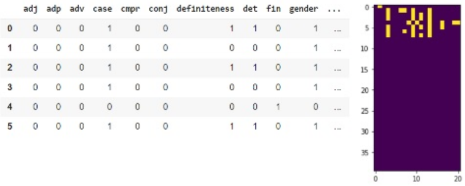
A Dataset Preparation Details

Truncation and padding. Since 95% of the sample sentences in the dataset had fewer than 40 words per sentence, we set the padding/truncation limit to 40, thus making each feature map to be a 40x40 pixel heat-map that encodes the labels for all the words in a sentence.

Heat maps. The entire source and target morphological data is represented as a 3-dimensional cuboidal heat map. Each sentence (entry) in the dataset is a single 2-D slice of the cuboid, the dimensions of which are [Padding Length] x [Number of Morpheme Categories]. The English-German pair, for example, has a sentence map dimension of [40 x 20]. Padding length is manually set based on the 95th percentile of sentence lengths across the dataset. Each row in the heat map would represent the morphological labels for a single word within the larger sentence. An example heat map excerpt is shown in figure 6.

B Details of the CNN model

A sequential convolutional network with 3 blocks of standard Convolution - Max Pool - Dropout - Batch Norm layers are used in the network. Relu activation and 'same' padding are used for all of the convolutional layers and a pool size of (2,2) is used for each MaxPooling2D layer. A fixed dropout of 0.2 is applied after each pooling layer in the three blocks. The output of the third block is up-sampled and flattened into a single-dimensional



	adj	adp	adv	case	cmpr	conj	definiteness	det	fin	gender	...
0	0	0	0	1	0	0	1	1	0	1	...
1	0	0	0	1	0	0	0	0	0	1	...
2	0	0	0	1	0	0	1	1	0	1	...
3	0	0	0	1	0	0	0	0	0	1	...
4	0	0	0	0	0	0	0	0	1	0	...
5	0	0	0	1	0	0	1	1	0	1	...

Figure 6: Heat map representation

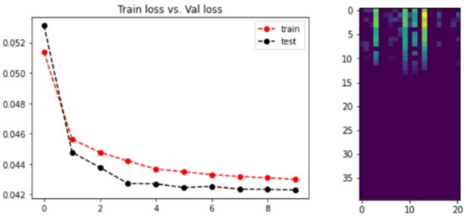


Figure 7: Sample output prediction

vector of length [categories x padding size]. A sigmoid activation is used in the final dense layer to facilitate the prediction of a probability score for every possible linguistic dimension of a word. The model is compiled with MSE as the loss function. A sample output prediction is shown in figure 7

C Fine-grained evaluations

Evaluation metrics of different granularities were explored to evaluate the model's performance. All the measures are computed at the heatmap level for each row (sentence) and averaged out over the entire dataset. We take standard accuracy, precision, and F1 scores for the flattened feature map vectors of the gold and predicted labels. Each feature map is originally of size `padding cut-off` times `number of linguistic dimensions` . Each unit of the predicted vector is independently compared with its corresponding gold label vector unit to evaluate the model output for different languages.

To get a more fine-grained sense of the model's performance, we explore two additional evaluation measures:

1. Proportion of missing labels

2. Proportion of extra labels

Tables 6 and 7 show fine-grained evaluations for both models. The missing label score represents the ratio of labels that are present in the gold gloss set but are absent in the model predictions. Similarly, the excess label score is the fraction of labels that have been wrongly predicted by the model.

D Variable training data experiments

Table 8 show results from experiments varying the amount of training data used.

Certain language families seem to demonstrate a significantly higher threshold for variance explainability based on dataset size. Spanish, French, and Italian (all romance languages) show massive jumps in accuracy from 20% to 40% training data but improve less drastically beyond the 60% training data mark. On the other hand, German and English show a rise in accuracy from 60% to 100% training data. Russian and Finnish demonstrate large jumps from 40% to 60% training data. Since the datasets' size was normalized before training, we might be able to conclude that these patterns are endemic to language families. For instance, it might be possible to conclude that the model requires significantly lesser training data to reach peak performance for romance languages as compared to Germanic languages. This generalization cannot be drawn from our small subset of languages and morphological tests, and therefore requires further investigation.

E Bi-directional LSTM for English Glossing

Figure 8 shows the model architecture for the Bi-directional LSTM that was used to gloss our source data and generate our source dataset for training purposes. The data was encoded with static w2v embeddings and the model was trained for 20 epochs (until convergence) on an English dataset containing UniMorph tags from the 2019 SIGMORPHON shared task referenced earlier. Model performance is detailed in table 2.

	CNN - Proportion of Missing Labels	**CNN - Proportion of Extra Labels**	**CNN - POS Accuracy**
Spanish	0.234	0.243	0.92
French	0.23	0.17	0.91
Basque	0.38	0.33	0.86
Italian	0.27	0.26	0.91
German	0.236	0.238	0.83
English	0.23	0.24	0.89
Turkish	0.39	0.32	0.8
Russian	0.36	0.33	0.84
Finnish	0.24	0.13	0.92

Table 6: CNN - Morpheme Tagging Scores, fine-grained evaluation

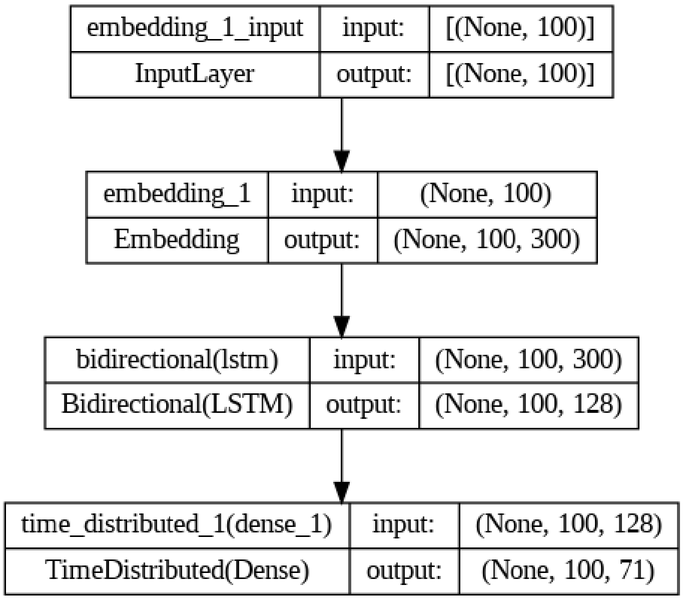

embedding_1_input | input: | [(None, 100)]
InputLayer | output: | [(None, 100)]
embedding_1 | input: | (None, 100)
Embedding | output: | (None, 100, 300)
bidirectional(lstm) | input: | (None, 100, 300)
Bidirectional(LSTM) | output: | (None, 100, 128)
time_distributed_1(dense_1) | input: | (None, 100, 128)
TimeDistributed(Dense) | output: | (None, 100, 71)

Figure 8: LSTM Model for English Text Glossing - Pseudo-parallel data generation

	BERT - Missing Labels	**BERT - Extra Labels**	**BERT - POS Accuracy**
Spanish	0.17	0.19	0.96
French	0.21	0.22	0.95
Basque	0.27	0.13	0.91
Italian	0.2	0.16	0.94
German	0.18	0.2	0.88
English	0.17	0.09	0.95
Turkish	0.25	0.13	0.87
Russian	0.24	0.28	0.88
Finnish	0.18	0.09	0.92

Table 7: BERT Morpheme Tagging Scores, fine-grained evaluation

Limited Train Data - F1 Score	**20%**	**40%**	**60%**	**80%**
Spanish	0.38	0.48	0.55	0.63
French	0.46	0.52	0.62	0.76
Basque	0.36	0.37	0.49	0.52
Italian	0.41	0.45	0.59	0.68
German	0.49	0.56	0.64	0.72
English	0.49	0.59	0.68	0.73
Turkish	0.33	0.38	0.52	0.58
Russian	0.39	0.4	0.64	0.71
Finnish	0.47	0.53	0.71	0.73

Table 8: Variable Training Data - Results

IGT Abbreviation	**UniMorph Abbreviation**
???	Unk
A1P	[’ABS’, ’1’, ’PL’]
A1S	[’ABS’, ’1’, ’SG’]
A2P	[’ABS’, ’2’, ’PL’]
A2S	[’ABS’, ’2’, ’SG’]
ADJ	ADJ
ADV	ADV
AFE	V
AFI	POS
AGT	AGFOC
AP	ANTIP
APLI	APPL
ART	ART, INDF
CAU	CAUS
CLAS	CLF
COM	PRF
COND	COND
CONJ	CONJ

Table 9: IGT to UniMorph Mappings